Tests for Gene Clusters Satisfying the Generalized Adjacency Criterion

Ximing Xu and David Sankoff

Department of Mathematics and Statistics,
University of Ottawa, Ottawa, Canada K1N 6N5
{xxu060,sankoff}@uottawa.ca

Abstract. We study a parametrized definition of gene clusters that permits control over the trade-off between increasing gene content versus conserving gene order within a cluster. This is based on the notion of generalized adjacency, which is the property shared by any two genes no farther apart, in the linear order of a chromosome, than a fixed threshold parameter θ. Then a cluster in two or more genomes is just a maximal set of markers, where in each genome these markers form a connected chain of generalized adjacencies. Since even pairs of randomly constructed genomes may have many generalized adjacency clusters in common, we study the statistical properties of generalized adjacency clusters under the null hypothesis that the markers are ordered completely randomly on the genomes. We derive expresions for the exact values of the expected number of clusters of a given size, for large and small values of the parameter. We discover through simulations that the trend from small to large clusters as a function of the parameter theta exhibits a "cut-off" phenomenon at or near $\sqrt{\theta}$ as genome size increases.

1 Introduction

Criteria for identifying common spatial groupings, such as synteny blocks or gene clusters, in two or more genomes entail a trade-off between increased content versus stricter order: if we require genes, motifs, segments, anchors or other elements (for which we will use the generic terms *markers*) of the group to be ordered identically within different genomes, so that we can have great confidence that these are genuine, evolutionarily conserved or functionally determined configurations, only relatively small groups are likely to satisfy this restrictive condition, so that the analysis misses large common genomic regions that only suffer small, perhaps insignificant, disruptions of common order. On the other hand, by allowing unrestricted scrambling of markers within the common groups (e.g., r-windows [2], max-gap [1] or "gene teams" [3]), we may be able to detect larger, more loosely structured groupings, but at least in the first analysis, must forgo accounting for local genome rearrangement, missing an important aspect of evolutionary history, and we relinquish the possibility of pinpointing extensive local conservation of order within the group.

We previously presented a parametrized definition of gene clusters that allows us to control the emphasis placed on conserved order within a cluster [6] and

A.L.C. Bazzan, M. Craven, and N.F. Martins (Eds.): BSB 2008, LNBI 5167, pp. 152–160, 2008.

hence to systematically explore the details of the content/order trade-off. The basis for this is the notion of generalized adjacency, which is the property shared by any two markers no farther apart, in the linear order of a chromosome, than a fixed threshold. Then a cluster in two or more genomes is just a maximal set of markers, where in each genome these markers form a connected chain of generalized adjacencies. Increasing the size of the threshold relaxes the degree of common ordering required, within a cluster, in the different genomes.

Nevertheless, for any fixed threshold, evolutionary rearrangements continue to disrupt the orders of markers on chromosome and will create, alter or destroy generalized adjacency clusters. Since even pairs of randomly constructed genomes may have some generalized adjacency clusters in common, the question arises of whether the number or size of these clusters is significantly larger than the random case. To answer such questions in this paper, we study the statistical properties of generalized adjacency clusters under the null hypothesis that the n markers are ordered completely randomly on the genomes (N.B. it suffices to randomize just one of the genomes, since relabeling markers can convert one of the genome to a canonical order, e.g., $1, 2, \ldots, n$, without changing the number, location and size of clusters).

2 Definitions

Our definition of generalized adjacency clusters is illustrated in Figure 1.

Definition 1. *Let V_X to be the set of markers in the genome X. These markers are partitioned among a number of total orders called* **chromosomes**. *For markers g and h in V_X on the same chromosome in X, let $gh \in E_X$ if the number*

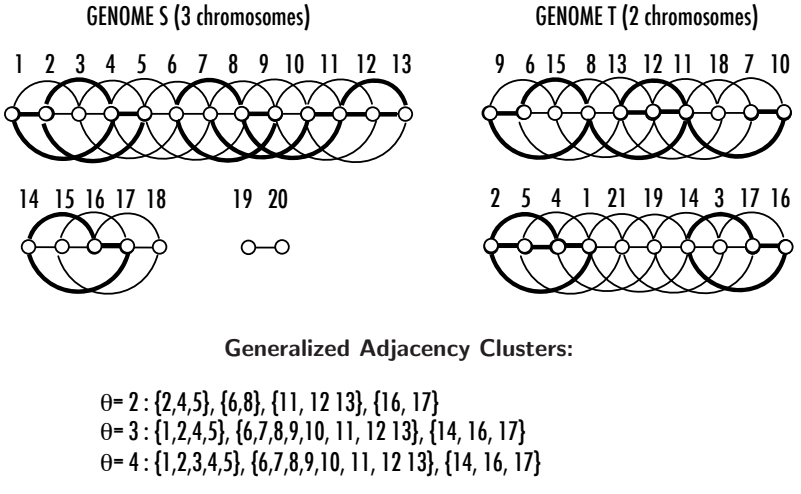

Fig. 1. Graphs constructed from two genomes using parameter $\theta = 3$. Thick edges determine generalized adjacency clusters. Clusters listed for $\theta = 2$ and $\theta = 4$ as well.

of genes intervening between g and h in X is less than θ, where $\theta \geq 1$ is a fixed **neighbourhood parameter**.

Consider the graphs $G_S = (V_S, E_S)$ and $G_T = (V_T, E_T)$ with a non-null set of vertices in common $V = V_S \cap V_T$. We say a subset of $C \subseteq V$ is a **generalized adjacency cluster** *if it consists of the vertices of a maximal connected subgraph of $G_{ST} = (V, E_S \cap E_T)$.*

This definition of clusters decomposes the markers in the two genomes into identical sets of disjoint generalized adjacency clusters of size greater or equal to 2, and possibly different sets of singletons belonging to no cluster, either because they are in V, but not in $E_S \cap E_T$, or because they are in $V_S \cup V_T \setminus V$. For simplicity, we do not attempt to deal with duplicate markers in this paper, and we also assume $V_S = V_T = V$. In practice, depending on the relative emphasis to be placed on order rearrangement versus marker insertion/deletion, we can delete all markers in $V_S \cup V_T \setminus V$ *before* calculating E_S and E_T, so as to exclude the effect of the markers unique to S or unique to T.

When $\theta = 1$, a cluster has exactly the same marker content and order (or reversed order) in both genomes. When $\theta = \infty$, the definition returns simply all the synteny sets, namely the sets of markers in common between two chromosomes, one in each genome.

3 The Number of Generalized Adjacencies in Common in Two Random Genomes

Each genome can be represented as a permutation of the first n positive integers. We denote by I the *reference genome* $1, 2, \ldots, n$ and by R the *random genome* sampled from all $n!$ possible genomes, each with probability of $\frac{1}{n!}$.

Let $n_2 = |E_I \cap E_R|$ denote the number of common edges, i.e. the number of the generalized adjacencies. For a random genome $R = r_1, r_2, \ldots, r_n$, if $r_h = i$, we define the *position* of i in R to be $g_i = h$. Then

$$|E_I \cap E_R| = |\{1 \leq i < j \leq n \mid j - i \leq \theta, |g_i - g_j| \leq \theta\}|.$$

Next we will study the probability distribution of n_2.

3.1 Large θ

A potential problem with generalized adjacency clustering, which it shares with other methods such as max-gap, is that beyond certain values of θ, instead of large clusters being statistically significant, the absence of such clusters becomes significant. We examine these cases first, before analyzing the more useful, smaller values of θ.

1. $\theta \geq n - 1$. In this case $n_2 = |E_I| = |E_R| = \binom{n}{2}$, so that $P[n_2 = \binom{n}{2}] = 1$.
2. $\theta = n - 2$
   (a) If $\{g_1, g_n\} = \{1, n\}$, probability $\frac{2}{n(n-1)}$, $n_2 = \binom{n}{2} - 1$,

(b) If $|(g_1, g_n) \cap \{1, n\}| < 2$, $n_2 = \binom{n}{2} - 2$.
Thus $P[n_2 = \binom{n}{2} - 1] = \frac{2}{n(n-1)}$ and $P[n_2 = \binom{n}{2} - 2] = \frac{(n-2)(n+1)}{n(n-1)}$

3. $\theta = n - k$ where k is a positive integer and smaller than $\frac{n}{2}$. In this case,

$$|E_I| = |E_R| = k(n-k) + \frac{(n-k)(n-k-1)}{2}.$$

Now, $|E_I \cap E_R| \geq |E_I| - \frac{k(k-1)}{2}$, because the number of the pairs (g_i, g_j), $i \neq j$ satisfying both $|i - j| \leq \theta$ and $|g_i - g_j| > \theta$ cannot be greater than $\frac{k(k-1)}{2}$. Then for k small relative to n,

$$n_2 \geq \binom{n}{2} - 2\binom{k}{2}$$

3.2 Small θ

$\theta = 1$. The definition of generalized adjacency reduces to the ordinary notion of adjacency. In this case the exact expression for the probability distribution of n_2 is known and its limiting distribution is Poisson with parameter 2 [4,5].

$\theta \geq 2$. We now present our main analytical results. We first examine the expected value $\mathbf{E}(n_2)$ of the number of adjacencies common to I and R.

Proposition. For $\theta \geq 1$,

$$\mathbf{E}(n_2) = 2\theta^2 - \frac{4n\theta^3 - \theta^2(1+\theta)^2}{2n(n-1)},$$

so that for a given θ

$$\lim_{n\to\infty} \mathbf{E}(n_2) = 2\theta^2$$

Proof. Counting the total number of edges in E_I, we have

$$|E_I| = (n-\theta)\theta + \sum_{i=1}^{\theta-1} i = n\theta - \binom{\theta+1}{2}$$

Each of these edges has probability

$$p = \frac{2(n-2)!}{n!} \sum_{i=1}^{\theta} (n-i)$$

of occurring in E_R. Thus

$$\begin{aligned} \mathbf{E}(n_2) &= |E_I| p \\ &= 2\theta^2 - \frac{4n\theta^3 - \theta^2(1+\theta)^2}{2n(n-1)}. \end{aligned}$$

□

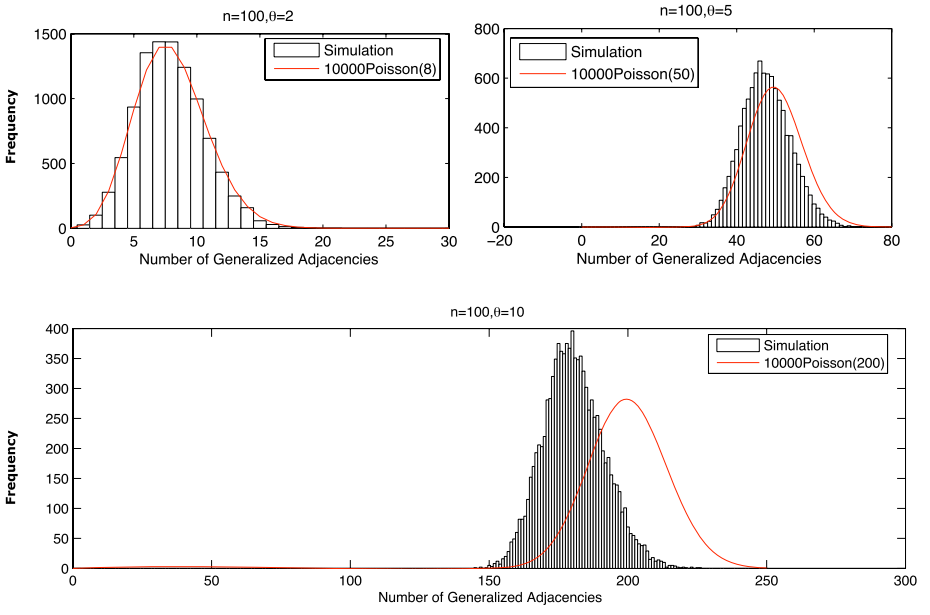

Fig. 2. Empirical distributions of the number of generalized adjacencies compared to the related Poisson distribution for $\theta = 2, 5$ and 10

We can say more about the limiting behaviour of n_2. Indeed, we may state (proof omitted):

Proposition. For $\theta \geq 1$, n_2 converges in distribution to a Poisson distribution with parameter $2\theta^2$.

We generated 10,000 random permutations on $1, \ldots, 100$ and calculated n_2 for various values of θ. In Figure 2, we compare the simulated distribution of n_2 (with means indistinguishable from $2\theta^2 - \frac{4n\theta^3 - \theta^2(1+\theta)^2}{2n(n-1)}$ in each case) to the Possion distribution with parameter $2\theta^2$, for $\theta = 2, 5$ and 10. For fixed n, the difference is larger as θ increases, though as n increases the Poisson is the limiting distribution.

4 Clusters of Larger Size

We use n_k to denote the number of connected components of size k in $E_I \cap E_R$, with no disjointness requirement or restriction against the component being contained in a larger cluster. We have already studied the distribution of n_2 in Section 3. We now consider the expectation of n_3. Extending the approach we used in the Proposition in Section 3.2, we can list all the connected components of size 3 in genome I and calculate the probability it is also in R. Adding all the probabilities together, we find

$$\mathbf{E}(n_3) = \frac{\theta^2}{n}(5\theta^2 - 2\theta - 1) + O(\frac{1}{n^2})$$

Similarly, with additional effort, we find that

$$\mathbf{E}(n_4) = \frac{\theta^2}{n^2}(\frac{124}{9}\theta^4 - \frac{95}{6}\theta^3 - \frac{8}{9}\theta^2 + \frac{29}{6}\theta + \frac{1}{9}) + O(\frac{1}{n^3})$$

but the number of different kinds of components of size 5 precludes extending our method, based on listing all possibilities, to n_5 and beyond.

4.1 Testing

Despite the fact that we have only partial results for n_k, we can still use standard statistical methods to test for the relatedness of two genomes or the significance of a generalized adjacency cluster, especially if $\mathbf{E}(n_4)$ is small.

5 The Maximum Size Generalized Adjacency Cluster

The ideal statistic to use to test the relatedness of genomes or to detect clusters would be the size of the largest cluster $k_{\max}$. While analytical techniques have not produced useful information about the distribution of $k_{\max}$, it is a straightforward matter to simulate random genomes and estimate this distribution empirically. Figure 3 shows the cumulative distribution functions for $k_{\max}$

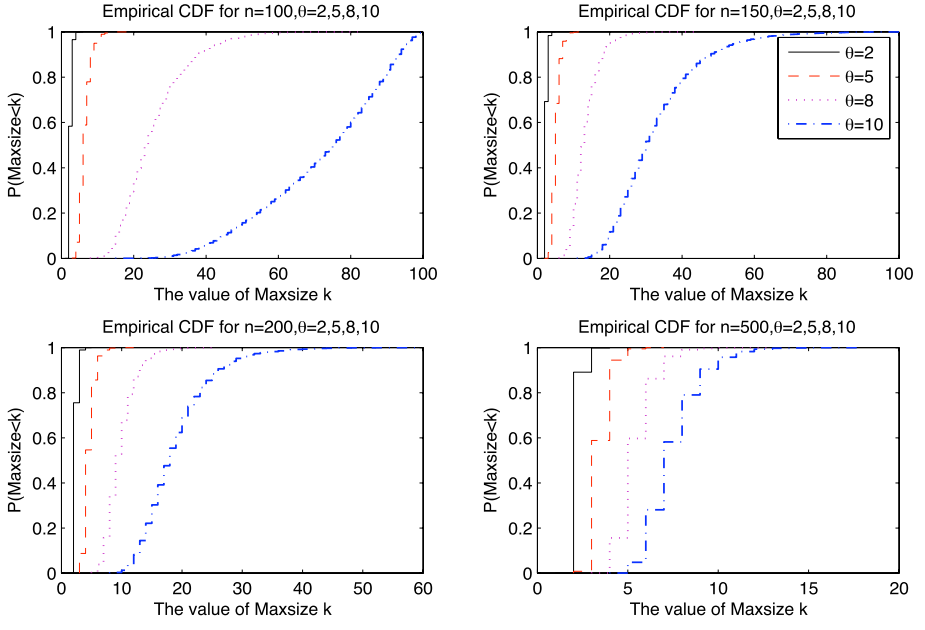

Fig. 3. Empirical cumulative distribution functions for $k_{\max}$ as a function of n and θ

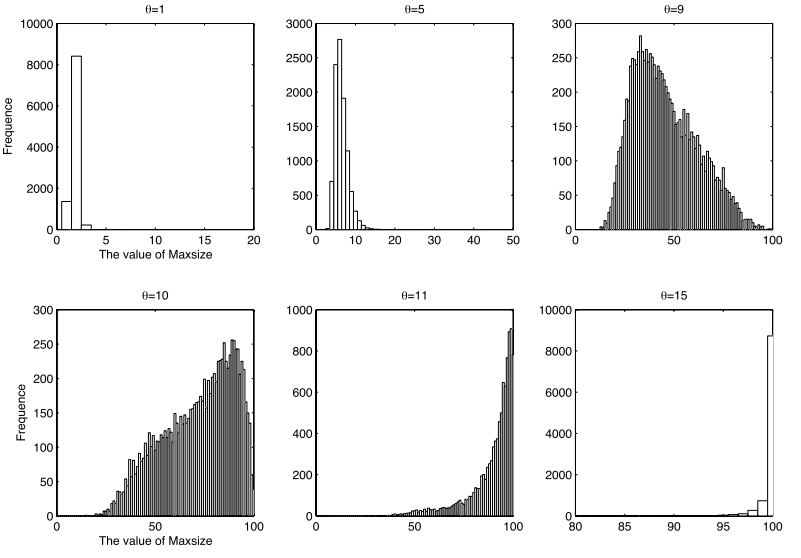

Fig. 4. Histograms for $k_{\max}$ when n=100

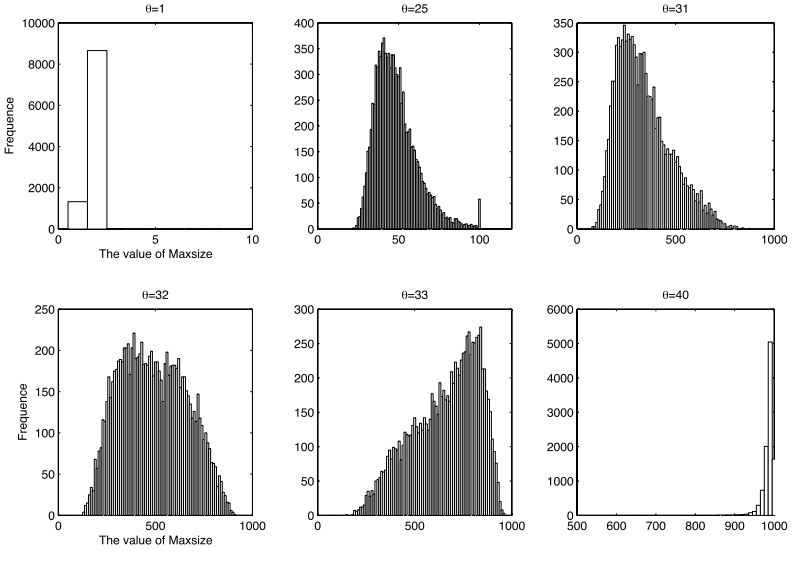

Fig. 5. Histograms for $k_{\max}$ when n=1000

as a function of n for a number of different θ. This kind of result can be directly used for testing.

Of particular interest is the dramatic change in the structure of the function between $\theta = 8$ and $\theta = 10$, when $n = 100$. Suddenly the mass of the distribution

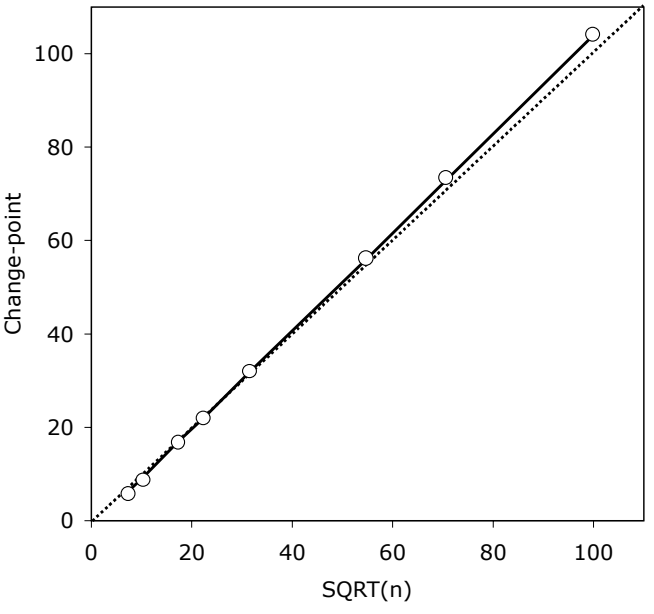

Fig. 6. Change-point for $k_{\max}$ distribution as a function of $\sqrt{n}$. Dotted diagonal represents exact square root of n.

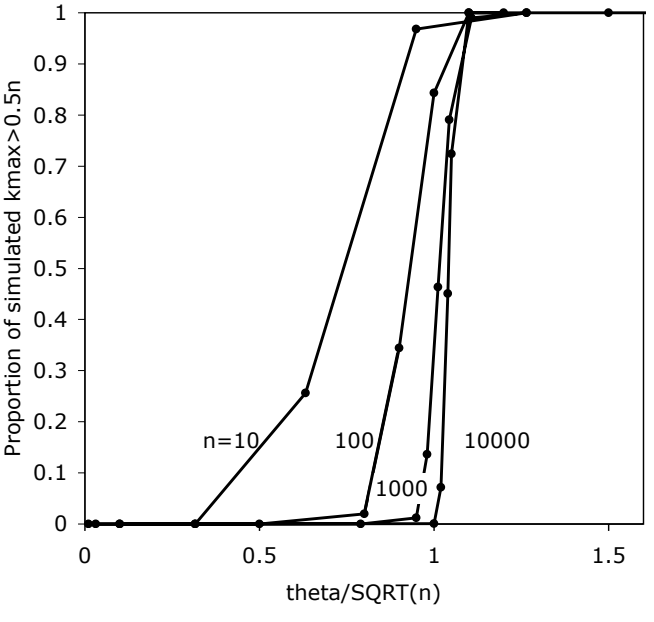

Fig. 7. Cut-off for maximum size cluster

shifts from values around 20 to values around 75. We investigated this phenomenon in more detail, as exemplified in Figures 4 and 5, each based on 10,000 pairs of random genomes. It is remarkable how quickly the distribution changes between $\theta = 9$ and $\theta = 10$ for $n = 100$, and between $\theta = 31$ and $\theta = 33$ for $n = 1000$. On the basis of 10,000 pairs of random genomes, we determined the mean change-point θ^* for a range of values of n, and in Figure 6 plotted these points against $\sqrt{n}$. This suggests that the change-point satisfies $\theta^* = \sqrt{n}$ or some similar relation.

To characterize the abruptness of the change around the change-point, we calculated how much of the probability mass falls to the right of $0.5n$, for each value of θ. Figure 7 shows that the change behaviour, in proportion to $\sqrt{n}$, tends to a sharp "cut-off" at or near $\theta = \sqrt{n}$.

6 Discussion

We have begun the investigation of statistics related to generalized adjacency clusters. The behaviour of the number of clusters for a given n and θ seems amenable to analytical investigation, as we have demonstrated with a number of new results. The distribution of $k_{\max}$, a tool for suggesting the biologically most interesting clusters, does not seem as accessible, but is easily simulated. Knowledge of the cut-off behaviour serves to delimit the region for meaningful tests to θ suitably less than $\sqrt{n}$.

Acknowledgments

Research supported in part by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) to DS. DS holds the Canada Research Chair in Mathematical Genomics.

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. Durand, D., Sankoff, D.: Tests for gene clustering. Journal of Computational Biology 10, 453–482 (2003)
3. 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)
4. Wolfowitz, J.: Note on runs of consecutive elements. Annals of Mathematical Statistics 15, 97–98 (1944)
5. Xu, W., Alain, B., Sankoff, D.: Poisson adjacency distributions in genome comparison: multichromosomal, circular, signed and unsigned cases. Bioinformatics 24 (2008)
6. Zhu, Q., Adam, Z., Choi, V., Sankoff, D.: Generalized gene adjacencies, graph bandwidth and clusters in yeast evolution. In: Mandoiu, I., Sunderraman, R., Zelikovsky, A. (eds.) ISBRA 2008. LNCS (LNBI), vol. 4983, pp. 134–145. Springer, Heidelberg (2008)