BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–8

Detecting Gene Clusters under Evolutionary Constraint in a Large Number of Genomes

Xu Ling*, Xin He

Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana IL 61801

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Spatial clusters of genes conserved across multiple genomes provide important clues to gene functions and evolution of genome organization. Existing methods of identifying these clusters often made restrictive assumptions, such as exact conservation of gene order, and relied on heuristic algorithms.

Results: We developed a very efficient algorithm based on a "gene teams" model that allows genes in the clusters to appear in different orders. This allows us to detect conserved gene clusters under flexible evolutionary constraints in a large number of genomes. Our statistical evaluation incorporates the evolutionary relationship among genomes, a key aspect that has been missing in most previous studies. We conducted a large scale analysis of 133 bacterial genomes. Our results confirm that our approach is an effective way of uncovering functionally related genes. The comparison with known operons and the analysis of the structural properties of our predicted clusters suggest that operons are an important source of constraint, but there are also other forces that determine evolution of gene order and arrangement. Using our method, we predicted functions of many poorly characterized genes in bacterial. The combined algorithmic and statistical methods we present here provide a rigorous framework for systematically studying evolutionary constraints of genomic contexts.

Availability: The software, data and the full results of this paper are available online at http://www.ews.uiuc.edu/~xuling/mcmusec.

Supplementary information: Additional methodological details and results are available in supplementary data.

Contact: xuling@uiuc.edu

1 INTRODUCTION

A fundamental problem in genomics is how genes are organized in the genomes, and what information is encoded in genomic contexts (Rogozin *et al.*, 2004). During evolution, the gene order is generally not well conserved because of the rapid rearrangement events that reshuffle genomes (Mushegian and Koonin, 1996; Huynen and Bork, 1998). On the other hand, functionally related genes may be constrained to remain close to each other due to natural selection, forming so called *conserved gene clusters* (Overbeek *et al.*, 1999). In prokaryotic genomes, these gene clusters are often manifested as *operons* (Jacob and Monod, 1961), genes typically involved in the same pathways and transcribed in single units. In eukaryotic genomes, operons are believed to be rare, but there are growing evidences that genome organization plays an important role in determining expression patterns of genes (Lawrence, 2002; Sproul *et al.*, 2005; Batada and Hurst, 2007; Ben-Shahar *et al.*, 2007).

Identification of conserved gene clusters across different genomes is important for several problems in comparative genomics (Rogozin *et al.*, 2004): (1) to predict operons or functional modules (Overbeek *et al.*, 1999; Kolesov *et al.*, 2001; Snel *et al.*, 2002; Dam *et al.*, 2007); (2) to annotate the uncharacterized genes, whose functions could be inferred from other genes belonging to the same clusters (Huynen *et al.*, 2000; Wolf *et al.*, 2001; Kim *et al.*, 2005); (3) to study genome organization and evolution (Wolf *et al.*, 2001; Lathe *et al.*, 2000; Yang and Sze, 2008); (4) to reconstruct species phylogeny from the information in the genomic organization (Kim *et al.*, 2005). In short, conservation of spatial organization of genes provide an important source of information that is orthogonal to primary sequences of genes and thus could be exploited to supplement our existing genomic analysis tools.

Some studies identify conserved gene clusters as gene strings with identical order across genomes (Overbeek *et al.*, 1999; Wolf *et al.*, 2001; Tamames, 2001). However, a gene cluster under functional constraint may still have experienced internal rearrangement events as long as these events do not disrupt the functioning of this group (Watanabe *et al.*, 1997; Itoh *et al.*, 1999; Lathe *et al.*, 2000). Therefore, the gene order may not be fully conserved. Other methods iteratively merge gene pairs that are close to each other in multiple genomes (Kolesov *et al.*, 2001; Rogozin *et al.*, 2002; Wu *et al.*, 2005; Zheng *et al.*, 2005; Price *et al.*, 2005). These procedures alleviate but do not fully solve the problem because under even minor rearrangements, not all gene pairs of a conserved cluster will remain adjacent. In addition, these studies often rely on heuristic algorithms that concatenate gene pairs by somewhat arbitrary criteria.

A series of papers developed the model of "gene teams", also called max-gap clusters, that relaxed these constraints of earlier methods (Bergeron *et al.*, 2002; He and Goldwasser, 2005; Kim *et al.*, 2005; Ling *et al.*, 2008). Under this model, a group of genes form a gene team if they remain spatially close in a set of genomes regardless of the internal order. The concept of a conserved gene cluster is precisely defined, thus it is possible to do formal algorithmic and statistical analysis, an advantage over heuristic methods. This model has been successfully applied to predict functional gene groups and operons (Luc *et al.*, 2003; He and Goldwasser, 2005) and study the evolution of protein

*to whom correspondence should be addressed

families (Pasek *et al.*, 2005). One major difficulty of the gene team model is its scalability with respect to the number of genomes. Apparently, using more genomes will increase the power of detecting natural selection on genomic organization, and provide a more comprehensive picture of genome evolution. Indeed, hundreds of prokarytoic and many eukaryotic genomes are currently available, and given the rapid progress of genome sequencing technology, we expect many more will come soon. It is thus crucial to have an efficient algorithm that can easily handle dozens of or even hundreds of genomes. Unfortunately, the computational complexity of earlier gene team algorithms is exponential to the number of genomes (Bergeron *et al.*, 2002; He and Goldwasser, 2005). This situation has been improved by the method of (Kim *et al.*, 2005), but as we will see later in our experiments, this method is still insufficient to process a large number of genomes under realistic parameter settings.

It is important to distinguish "phylogenetic inertia", conservation of gene clustering due to inadequate time of reshuffling genomes, from true evolutionary constraint because only the later provides functional information (Rogozin *et al.*, 2004). Thus, a conservation measure taking into account this distinction is important. The number of genomes where a gene group remains clustered has often been used as the criterion for defining a conserved cluster (Snel *et al.*, 2002; Tamames, 2001; Wolf *et al.*, 2001; Rogozin *et al.*, 2002; Kim *et al.*, 2005), but this is not a very good indicator of evolutionary constraint since the genomes are not uniformly sampled for sequencing. For example, two gene clusters conserved in equal numbers of genomes may represent very different functional constraint, if one of them is only conserved in closely related species. The current statistical tests for the gene team model also have limitations. The tests proposed in He *et al.* (He and Goldwasser, 2005) and Hoberman *et al.* (Hoberman *et al.*, 2005) essentially assume that the genomes being compared are very distant so that any shared organization due to common ancestry has been lost. As such, the tests are not applicable to a larger number of genomes when there is no clear-cut distance.

We previously adopted the gene team model for two-species comparison (He and Goldwasser, 2005; Ling *et al.*, 2008). In the current work, we extended our earlier algorithms to analysis of an arbitrary number of genomes. The greatly improved computational power allows us to do a large-scale discovery of conserved gene clusters with flexible evolutionary constraints. We borrowed the notion of branch length score (BLS) from the study of regulatory elements (Kheradpour *et al.*, 2007) to quantify the evolutionary constraint of gene clusters, taking into account the phylogenetic tree structure of species. Our computational approach is applied to more than 100 bacterial genomes. We predicted a large set of conserved gene clusters and made some interesting observations regarding the function and evolution of these clusters.

2 METHODS

2.1 Defining conserved gene clusters

We first note that the basic unit in our framework, for which clustering is defined, may be a gene, a domain (Pasek *et al.*, 2005), a homologous sequence anchor (Ling *et al.*, 2008), or even a functional sequence element such as a transcription factor binding site. This generality is one advantage of our framework because the same algorithm can be applied to all these cases without change. In the following we assume that basic units are genes. We also assume that homology of genes has already been given, yet the algorithm itself is independent of the method used to infer homology.

It is informative to examine the problem of other existing methods to understand why a different approach is needed. Many methods for detecting conserved gene clusters start with search of gene pairs that are close in multiple species, then merge those pairs (Kolesov *et al.*, 2001; Rogozin *et al.*, 2002; Wu *et al.*, 2005; Zheng *et al.*, 2005; Price *et al.*, 2005). The limitation of this general procedure has been discussed in details in (Hoberman *et al.*, 2005; Bergeron *et al.*, 2002; Ling *et al.*, 2008). The main problem is that it tolerates very few "conservative rearrangement" events within a cluster even if the genes of this cluster overall remain close to each other. As earlier researchers pointed out, such events are not uncommon (Lathe *et al.*, 2000; Rogozin *et al.*, 2004).

Our definition is based on the notion of gene teams (Bergeron *et al.*, 2002). We term a gene and all its homologs, including its orthologs and paralogs, as a *homology family*. A genome is then defined as an ordered sequence of genes where a gene is associated with the homology family it belongs to. The distance between two genes can be defined by either the number of intervening genes or the number of intergenic nucleotides. A gene team defines a set of genes that remain spatially close in a given set of genomes regardless of orders. Specifically, given two parameters $minsize$ and $maxgap$ and a set of genomes, we define a set of genes as a "complete max-gap cluster" if the number of genes in this set is no less than $minsize$, and in each of the input genomes, the distance between any pair of adjacent genes in this set is always no more than $maxgap$. Also note that in this definition, genes are allowed to be in either DNA strand. A toy example for pairwise comparison is demonstrated in Figure 1.

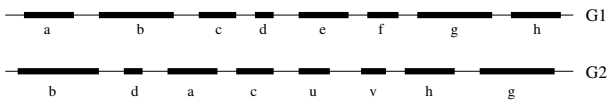


Fig. 1. Two example genomes $G_1 = \{abcdefgh\}$ and $G_2 = \{bdacuvhg\}$, where $a, b, \ldots v$ are genes. Measured by gene insertions, the distance between a and b on G_1 is 0, but 1 on G_2. The maximal distance between any adjacent pair of members of the group of 4 genes: $\{a, b, c, d\}$ is 0 in both G_1 and G_2. Given the threshold $maxgap = 0$ and $minsize = 2$, the gene set $\{a, b, c, d\}$ qualifies as a complete max-gap gene cluster with respect to genome G_1 and G_2, while its subsets (*e.g.*, $\{a, b\}$, $\{b, c\}$, $\{c, d\}$) are not.

When performing a large scale study on many genomes, we relax the conservation constraint so that a cluster does not have to be conserved in all input genomes. Given the parameter $minsupp$, we will call a set of genes a "frequent max-gap cluster" if it occurs in at least $minsupp$ genomes as a complete max-gap cluster. For convenience, we will also call a genome where a cluster actually appears subject to the proximity constraint, as a supporting genome of this cluster. Given N genomes, the goal is to find all frequent max-gap gene clusters with respect to given parameters $maxgap$, $minsize$ and $minsupp$.

2.2 The MCMuSeC algorithm

MCMuSeC, **M**ax-gap **C**lusters by **Mu**ltiple **Se**quence **C**omparison, is our algorithmic framework that computes conserved gene clusters through N-genomes comparison in two phases. The work by Kim *et al.* (Kim *et al.*, 2005) shares a similar goal. Their algorithm is based on the *Apriori* heuristic well-known in the field of data mining (Agrawal and Srikant, 1994). In addition to applying the same idea, our algorithm employs another optimization strategy, which turns out to be crucial for the effiency. The details of our algorithm are described below, with a concrete example at Figure 2.

We start with the case of identifying complete max-gap clusters that appear in all $M = minsupp$ genomes, given the parameters $maxgap$ and $minsize$. For simplicity, we will call a group of genes in a genome G as a gene *subset*. We denote the genomes $G_1, G_2, \ldots, G_M$. At each step, the algorithm takes subsets $S_1, S_2, \ldots, S_M$ ($S_m \subseteq G_m$) as input, and recursively performs the *decompose* and *filter* procedures. Initially, S_m is set to be the genome G_m for $m = 1, \ldots, M$. The *decompose* procedure is essentially an extension of the HomologyTeam (He and Goldwasser, 2005) approach. Specifically, it consists of four steps:

1. Scan M subsets $S_1, S_2, \ldots, S_M$, and remove the genes which do not have a homolog in all subsets;
2. Sort genes on each subset according to their positions on the chromosome;
3. Break the subset S_m ($m = 1, \ldots, M$) into several smaller subsets at the positions where the distance between the flanking neighboring genes exceeds the distance threshold $maxgap$. Apparently, these gene pairs certainly can not belong to the same max-gap cluster;
4. Remove produced subsets whose size is less than $minsize$.

Let S'_m denote the remaining genes on S_m (after uncommon genes being removed in step 1). If none of the input subsets S_m was decomposed, the algorithm outputs $S'_1, \ldots, S'_M$ as a complete max-gap cluster. If all genes on S_m are removed for some m, the M input subsets certainly do not contain a valid gene cluster, and the procedure will terminate with no output. Otherwise, any combination of the decomposed smaller subsets will be submitted to the *filter* procedure.

Next, we use the *filter* procedure to quickly eliminate the false combinations which will not generate any valid gene clusters. Suppose the subset S_m is decomposed into D_m smaller subsets $S_m^1, S_m^2, \ldots, S_m^{D_m}$ ($m = 1, \ldots, M$). A *combination* $(S_1^{d_1}, S_2^{d_2}, \ldots, S_M^{d_M})$ (where $S_m^{d_m} \subseteq S_m$, $m = 1, \ldots, M$) is defined by a composition of M smaller subsets, where the m^{th} subset comes from S_m. A brute-force approach would enumerate all such combinations, hence the complexity is $\prod_{m=1}^{M} D_m$. We are able to quickly identify the promising combinations by leveraging the concept of *signatures*. Basically, a signature is defined by a certain number of genes, which might belong to the final resulted cluster. It is not hard to see that if there exists a valid gene cluster for $(S_1, S_2, \ldots, S_M)$, then there must exist at least $minsize$ number of gene members conserved in all M subsets. Let's refer the set of t-gene composition from S_m to *the size-t signatures* of S_m. The intuition behind the filtering idea is that, if we enumerate all size-t signatures (*i.e.*, a set of t genes, $t \leq minsize$) for each S_m and find that a combination has no common size-t signatures, we can immediately discard this combination since it definitely will not produce any valid gene clusters.

We now discuss how to effectively generate size-t signatures and use them for pruning. The algorithm computes a set of size-t signatures $\text{SIG}(S_m^{d_m}) = \{\text{sig}_1, \text{sig}_2, \ldots, \text{sig}_{\text{NUM}}\}$ for each obtained subset $S_m^{d_m}$ ($d_m = 1, 2, \ldots, D_m$), where each sig_i consists of t genes from $S_m^{d_m}$. Note, the value of NUM (*i.e.*, the number of size-t signatures) depends on t, and we denote it as $\text{NUM}(t)$ thereafter. To associate each $S_m^{d_m}$ with its signatures, we maintain a M-way lists E_s for each signature s, where the m^{th} list $E_s[m]$ stores the subsets $S_m^{d_m}$ which generate s. After signatures from each subset are all generated, we check each E_s to identify combinations of subsets with common signatures. That is, for each E_s the algorithm outputs combinations $(S_1^{d_1}, S_2^{d_2}, \ldots, S_M^{d_M})$, where $S_m^{d_m} \in E_s[m]$, then recursively apply the *Decompose-Filter* algorithm on the generated combinations.

We further observe that it is not necessary to generate all t-gene compositions as size-t signatures. In Supplementary Materials, we show how to compute and minimize the number of size-t signatures with theoretical proofs. This optimization can essentially bring two benefits: first, the complexity in real computation is reduced; and second, more combinations are pruned.

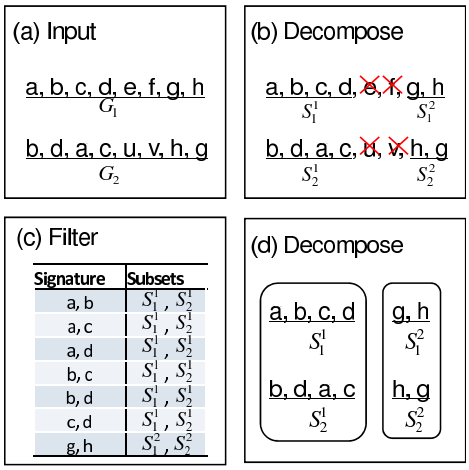


Fig. 2. The Decompose-Filter procedure of MCMuSeC algorithm. The example is taken from Figure 1 with parameters $maxgap = 0$ and $minsize = 2$. (a) The input genomes. (b) The *decompose* procedure first matches genes between G_1 and G_2, and removes non-homologous genes e, f from G_1 and u, v from G_2 respectively. It identifies two subsets on G_1: $S_1^1 = \{a, b, c, d\}$ and $S_1^2 = \{g, h\}$, and two subsets on G_2: $S_2^1 = \{b, d, a, c\}$ and $S_2^2 = \{h, g\}$. That results in total four possible combinations of subsets: (S_1^1, S_2^1), (S_1^1, S_2^2), (S_1^2, S_2^1), (S_1^2, S_2^2). (c) Suppose $t = 2$, the *filter* procedure first generates signatures for all 4 subsets. That is, $SIG(S_1^1) = \{(a, b), (a, c), (a, d), (b, c), (b, d), (c, d)\}$, $SIG(S_1^2) = \{(g, h)\}$, $SIG(S_2^1) = \{(b, d), (a, b), (b, c), (a, d), (c, d), (a, c)\}$ and $SIG(S_2^2) = \{(g, h)\}$. Then, it groups the subsets by signatures as shown in the table. The signatures shared by two smaller subsets which come from different original subsets indicate the subset combination which might lead to valid clusters. (d) The filter procedure finally submits (S_1^1, S_2^1) and (S_1^2, S_2^2) to the next round of *Decompose-Filter* iteration. As neither (S_1^1, S_2^1) nor (S_1^2, S_2^2) is decomposable, the algorithm will output two gene clusters (S_1^1, S_2^1) and (S_1^2, S_2^2) as the final results.

To generalize to the case where the number of genomes, $N \geq minsupp$, we exploit the *Apriori* heuristic (Han *et al.*, 2004; Kim *et al.*, 2005) because the brute-force approach that examines all combinations of $minsupp$-genomes is too expensive. We start with $M = 2$ genomes, and progressively increase M by including one more genome. A $(M+1)$-genome combination is examined only if the M-genome combination generates some valid gene clusters. At any stage where no max-gap cluster is generated, we stop adding genomes and the algorithm terminates. Notice, when we compute the $(M + 1)$-genome combination (by including a new genome G to the M-genome combination), we do not submit M original genomes together with the new genome G to *Decompose-Filter*. Instead, genes that do not occur in any gene clusters in the M-genome combination (according to the *Apriori* computation on the M-genome) are removed. This strategy enables the *decompose* procedure to generate much finer subsets, so that the *filter* procedure becomes more effective in pruning.

2.3 Quantifying the evolutionary constraints of gene clusters

In general, we do not know the correct value of $minsupp$ and a more fundamental difficulty is that the number of genomes where a cluster appears may not be a good measure of evolutionary constraint. This issue is

recognized by Zheng *et al.* (Zheng *et al.*, 2005), who developed a statistical measure of conservation for gene pairs, taking into account the phylogenetic relationship among species. Their measure is similar to *Branch Length Score* (*BLS*), initially proposed for regulatory motif prediction (Kheradpour *et al.*, 2007). The intuition of BLS is that the longer evolutionary time a cluster is conserved, the more likely it is under constraint. We adopted BLS for our evaluation of evolutionary constraints on gene clusters. We note that despite the similarity of the statistical measure, the work by Zheng *et al.* (Zheng *et al.*, 2005) is based on pairs of genes and thus suffers from the problem of allowing few changes inside clusters, as discussed before.

The BLS value of a conserved gene cluster is defined as the total length of all branches of the phylogenetic tree where the cluster is conserved. For each putative gene cluster, we evaluate the statistical significance of its BLS against the randomly sampled population of gene clusters, which serves as our null hypothesis. Specifically, for a gene cluster of size k, we randomly sample 1000 size-k clusters, under the same $maxgap$ constraint, from one supporting genome where the cluster appears. Then we compute the BLS value for each random cluster. This process is repeated for each supporting genome, and the pooled collection of all BLS values defines the null distribution. The proportion of BLS values which are no less than the BLS value of the predicted cluster will be used as the p-value.

One problem of the above approach is multiple hypothesis testing. Each cluster is tested for its statistical significance, but many clusters are tested simultaneously, so some form of correction of multiple hypothesis testing will be needed. This is particularly a problem for large clusters as the number of possible large clusters from a genome is large. We adopted some form of Bonferroni correction. Under this scheme, the p-value cutoff should be equal to 0.05 divided by the number of hypothesis tested. We notice that under $maxgap = 2$ (the default setting), to form a size-k cluster from a cluster of size $(k-1)$, a new gene can be added in 3 possible positions with distance to the nearest gene equal to 0, 1 or 2 respectively, therefore, the number of clusters (also hypothesis to be tested) of size k is three times the number of clusters of size $(k-1)$. So our strategy is: for small clusters (size less than or equal to 5) we use p-value 0.05 as cutoff, and when the cluster size is increased by one, we will reduce the p-value threshold by $1/3$.

2.4 Genomes and the phylogenetic tree

We used 133 bacterial genomes from NCBI's Genome Assembly/Annotation Projects FTP site (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/), and the phylogenetic tree from (Ciccarelli *et al.*, 2006). The length of a branch is measured by the average number of nucleotide substitutions in that branch.

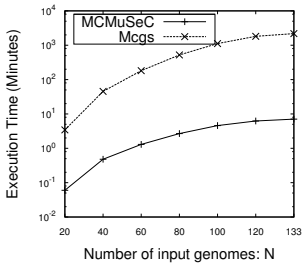
2.5 Extracting conserved gene clusters at *E. coli*

The results from the MCMuSeC algorithm may contain overlapped clusters. For example, three genes A, B and C may form a conserved cluster in some genomes, but A, B and another gene D may also form a conserved cluster in another set of genomes. To address this redundancy, we construct a set of disjoint clusters from the model organism *E. coli*. From all predicted clusters from the analysis of 133 genomes that are also statistically significant, we first sorted them by their significance. In the order of decreased significance, we checked each predicted cluster. If the current cluster does not overlap with any previously identified *E. coli* clusters, we will add it to the current set of *E. coli* clusters. Otherwise, we will discard this cluster and continue on the next one.

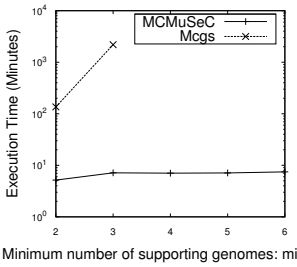
3 RESULTS

3.1 Discovery of conserved gene clusters in more than 100 bacterial genomes

We applied our algorithm in 133 bacterial genomes. The homologous relationship of genes is based on COG (Clusters of Orthologous Groups) annotation (Tatusov *et al.*, 2003). That is, all genes annotated with the same COG would be considered as homologs. We first compared the computational performance of our algorithm (MCMuSeC) with the algorithm, Mcgs, by (Kim *et al.*, 2005). The Mcgs algorithm is also based on the gene team formulation, but applies different ideas for speeding up computation. The earlier gene team algorithms, GeneTeams (Bergeron *et al.*, 2002) and HomologyTeams (He and Goldwasser, 2005), have running time exponential to the number of genomes, thus cannot even be tested under our setting. Experiments were conducted on a Linux Server with 2.2GHz AMD Opeteron Processor and 32GB RAM.



(a)



(b)

Fig. 3. Computational performance of MCMuSeC and Mcgs. (a) Execution time vs. the number of input genomes N, with $maxgap = 200$ base pairs, $minsize = 2$, $minsupp = 3$; (b) Execution time vs. $minsupp$, with $N = 100$, $maxgap = 200$ base pairs, $minsize = 2$.

Figure 3 shows the performance of the two algorithms. At the small value of $minsupp$ (3 in this case), both methods seem to scale well with respect to the number of input genomes N, when N is relatively small. However the computation time of Mcgs increases much faster than MCMuSeC when N gets larger (Figure 3(a)). For instance, at $N = 133$, Mcgs took more than 36 hours, while our method only needed 7 minutes. At larger values of $minsupp$, as shown in Figure 3(b), two methods perform dramatically different. For $minsupp \geq 4$, Mcgs was even not able to finish the computation in 7 days, while MCMuSeC generated results in less than 8 minutes. We also observed similar patterns under different settings of $maxgap$ (data not shown). Our algorithm applies two techniques for optimization: the filter procedure, and the *Apriori* heuristic, while Mcgs only exploit the *Apriori* idea. Our results suggest the combination of the two optimization techniques is extremely effective in reducing the computational cost.

We note that we only compare the computational efficiency of MCMuSeC with Mcgs, but not accuracy of the two programs, because both are based on the same gene team formulation and will generate identical results with the same input and parameter

settings. On the other hand, the results from MCMuSeC will be subject to further statistical evaluation and the insignificant clusters will be filtered out. Such statistical procedure is not available in Mcgs.

We identified gene clusters in $N = 133$ genomes under the setting: $minsize = 2$, $maxgap = 2$ gene insertions and $minsupp = 3$. This parameter setting is based, to some extent, on early empirical studies (Wolf *et al.*, 2001; Kim *et al.*, 2005). Most importantly, unlike these earlier studies, we will use the statistical evaluation procedure to choose the truly functional clusters, rather than depend on knowing the accurate values of these parameters. We found a total of 32,825 gene clusters, among which we further extracted 370 statistically significantly ($p < 0.05$) conserved clusters occurring at *E. coli* K12. We used these 370 clusters in all the subsequent experiments except the last one.

3.2 Functional characterization of conserved gene clusters in *E. coli*

We would suspect that the genes of a conserved cluster are functionally related. We used the 22 functional categories of COGs in NCBI (http://www.ncbi.nlm.nih.gov/COG/grace/fiew.cgi), as our annotation of the corresponding genes. We measure the functional coherence of a gene cluster as the p-value of the hypergeometric test that assesses enrichment of a functional category in the genes of that cluster. Figure 4(a) displays the distribution of this measure from all clusters. For comparison, we also drew the distribution of a set of size-matched gene clusters randomly sampled from the *E. coli* genome. Our identified gene clusters are much more likely to share the same function than the random clusters. For example, there are 223 out of 370 clusters enriched with certain functional category with p-value ≤ 0.01. These results strongly suggest that our method is indeed an effective way of recovering functionally related genes.

We then compared our identified gene clusters with 650 experimentally validated multi-gene operons in *E. coli* from RegulonDB (Salgado *et al.*, 2006). We first tested to what extent clustering of known operons is conserved. We extracted the occurrences of the known operons in the above 133 genomes, and applied the same statistical assessment as we did before. Among 650 multi-gene operons, 192 (30%) operons are conserved with p-value ≤ 0.05. This lack of strong conservation of operons is consistent with earlier studies using a much smaller set of genomes (Itoh *et al.*, 1999).

Next, for each of the 370 conserved clusters in *E. coli*, we define its "operon-coverage" as the number of known operons that overlap with this cluster. We plotted the histogram of "operon-coverage" in Figure 4(b). 173 gene clusters match exactly to single known operons, but there are also a substantial portion of conserved clusters that cover two or more operons. This pattern is consistent with the notion of "uber-operons" spanning possibly multiple adjacent operons, proposed earlier by other researchers (Lathe *et al.*, 2000). There are 18 gene clusters, listed in Supplementary Materials Table 1, that are not covered by any of the known multi-gene operons. As an indication that these clusters are likely under functional constraints, we found that nine of them had a consensus category (not including the category R). We examined one of the remaining clusters consisting of five genes *yraL*, *yraM*, *yraN*, *diaA* and *yraP*. This cluster is conserved across 15 genomes with BLS 1.343. The exact arrangement of this cluster, including gene order and orientation, is identical among all its occurrences. This high level of conservation strongly suggests that this cluster is under purifying selection for its organization. Among the five genes, *yraM* and *yraP* are both lipoproteins involved in outer membrane integrity and essential for cell growth (Akerley *et al.*, 2002; Onufryk *et al.*, 2005). The gene *diaA* is known to be important for DNA replication (Keyamura *et al.*, 2007), and the gene *yraN* is annotated with COG 0792, "predicted endonuclease distantly related to archaeal Holliday junction resolvase". It is thus possible that this cluster represents related genes with a function in cell division. Interestingly, the first gene in the cluster, *yraL*, has a different orientation than other genes, thus, this cluster is apparently not a single unit of transcription.

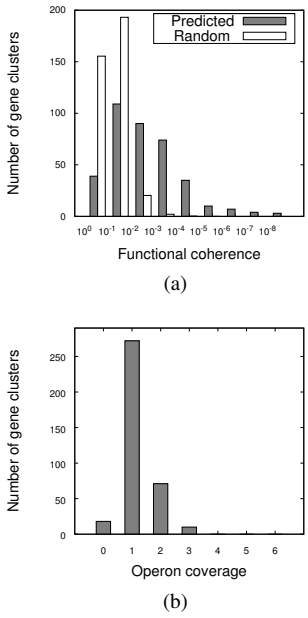


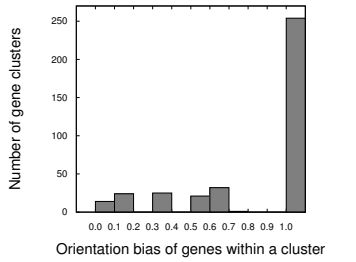
Fig. 4. Functions of conserved gene clusters in *E. coli*. (a) Statistical significance of the cluster's consensus category, defined by the p-value of hypergeometric test, for predicted clusters (filled bars) and random clusters (open bars); (b) Operon coverage of conserved gene clusters in *E. coli*. Operon coverage of a gene cluster is defined as the number of known operons from RegulonDB that overlap with this cluster.

Overall, our results demonstrate that operons play an important but not exclusive role of determining the evolution of gene arrangement. The fact that we were able to predict novel clusters with possibly related functions supports the utility of our method to discover new relationship among known genes. It is important to note that *E. coli* is an extensively studied model organism, thus it is not surprising that relatively little new biology can be learned. When studying genomes of less characterized organisms, we expect that our program will be much more useful.

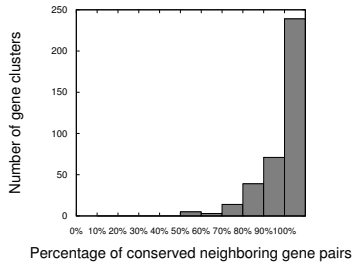
3.3 Organizational features of conserved gene clusters in *E. coli*

We next sought to characterize the organizational features of the conserved gene clusters: how often are they transcribed in the same direction? And how often is the gene order changed during evolution? For each cluster, we first calculate the percentage of

genes transcribed in each of the two strands, then we define the orientation bias of this cluster as the difference of the two percentages (the absolute value). It is 0 if there are equal numbers of genes in both strands and 1 if all genes are in the same strand. We found 254 out of 370 (68%) gene clusters within which all genes showed the same orientation (Figure 5(a)), suggesting a strong effect of gene co-regulation. Still, this percentage is much smaller than that based solely on operons. In a recent study (Yang and Sze, 2008), it was found that in 92% of time, the orthologous genes of an *E. coli* operon had the same orientation in other genomes.



(a)



(b)

Fig. 5. Organizational features of conserved gene clusters in *E. coli*. (a) Distribution of orientation bias of gene clusters, see text for definition of orientation bias; (b) Distribution of percentage of conserved neighboring pairs of gene clusters.

We also studied how gene orders are preserved across genomes. For each conserved gene cluster, we computed the percentage of adjacent gene pairs (in *E. coli*) that remain adjacent in another genome where the cluster occurs. Based on this measure, it is clear that an overwhelmingly large proportion of gene order is preserved (Figure 5(b)). Furthermore, among all 370 clusters, 239 (64%) have fully conserved spatial arrangement. One previous study on the rearrangements of operons (Yang and Sze, 2008) showed 84.8% of all operons had perfectly conserved arrangements. Thus, the two studies were largely consistent in proving the tendency of gene order conservation. On the other hand, the extent of conservation is significantly lower in our estimation.

We believe the analysis we presented here avoids the ascertainment bias problem created by using only known operons (Yang and Sze, 2008), and is thus more relevant to the study of the general patterns of genome evolution. Overall, our results support that conserved gene clusters tend to have uniform (in terms of orientation) and highly conserved organization. Note that in defining and identifying the conserved clusters, we put no requirement on gene orientations, and no constraint on the internal order of genes within a cluster, therefore our observation does not come from the putative bias from our definition. On the other hand, it is clear that genes in a cluster are not always on the same strand, and small changes of gene order are tolerated in evolution. These findings support our gene team model, which allows more flexible structure than earlier ones, and confirm that there are forces other than operons that influence the evolution of genome organization.

3.4 Prediction of functional categories of uncharacterized COGs

One very important application of genomic context analysis is to annotate the uncharacterized genes, whose functions can be inferred from other genes belonging to the same cluster (Overbeek *et al.*, 1999; Wolf *et al.*, 2001; Kim *et al.*, 2005). From the highly conserved clusters identified from all 133 bacterial genomes (not just those mapped to *E. coli*), we extracted 88 clusters in which more than 70% genes shared the same category code and at least one gene was annotated with COG in category R (General function prediction only) or S (Function unknown). For each cluster, the most frequent category code was used to predict the poorly characterized COGs. Table 1 lists 10 most conserved gene clusters (ranked by p-value of the BLSs) among the 88 predictions (see Supplementary Materials for the full list). A previous study of bacterial genomes also predicted functions of COGs based on genomic context conservation (Wolf *et al.*, 2001). Among their predictions, only 14 overlapped with ours, so we had 69 completely novel predictions. Note that there are two important differences between our study and that by Wolf *et al.* (Wolf *et al.*, 2001): first, gene order of a cluster must be fully conserved in their study; second, they did not provide a rigorous statistical evaluation of their predictions. In fact, any "conserved gene string" that appeared in more than three genomes, a somewhat arbitrary cutoff, was used for their prediction.

4 DISCUSSION

We developed a very efficient algorithm of identifying conserved gene clusters in a large number of genomes. Compared with other methods in this area, our scheme would allow one to detect and characterize genes clusters that are under evolutionary constraint but still have undergone some minor rearrangements. As earlier researchers pointed out (Lathe *et al.*, 2000; Rogozin *et al.*, 2004) and our study showed, such "conservative rearrangements" are not uncommon, and may be important in fine-tuning the expression pattern of the genes in different species (Price *et al.*, 2006). Our statistical method allows one to detect the gene clusters truly constrained by selection pressures from those with shared organization due to common ancestry. It has been recognized that the distinction between the two causes is crucial for utilizing the power of conserved genomic organization to make functional predictions (Rogozin *et al.*, 2004), yet most existing methods did not treat the problem in a rigorous way. We also point out the statistical method is particularly important for the gene team model we and other researchers have used. By relaxing the condition of clustering, this model may increase the false positive predictions comparing with earlier methods that require exact conservation of gene order. Our statistical test guards against this possibility by filtering out

Table 1. Fucntional predictions of uncharacterized COGs from 10 most conserved gene clusters

COG	Current COG functional annotation	Predicted function category	BLS	[a]p-value	[b]Evidence
3383	Uncharacterized anaerobic dehydrogenase	C: Energy production and conversion	1.713	0.0	11/12
1422	Predicted membrane protein	J: Translation, ribosomal structure and biogenesis	2.336	0.0	8/9
[c]2001	Uncharacterized protein conserved in bacteria	M: Cell wall/membrane/envelope biogenesis	4.911	0.0	8/9
4674	Uncharacterized ABC-type transport system, ATPase component	E: Amino acid transport and metabolism	4.578	$1.9 \cdot 10^{-3}$	4/5
1556	Uncharacterized conserved protein	C: Energy production and conversion	4.325	$3.0 \cdot 10^{-3}$	4/5
[c]2106	Uncharacterized conserved protein	J: Translation, ribosomal structure and biogenesis	1.707	$3.3 \cdot 10^{-3}$	4/5
0316	Uncharacterized conserved protein	C: Energy production and conversion	1.053	$3.6 \cdot 10^{-3}$	5/6
1738	Uncharacterized conserved protein	H: Coenzyme transport and metabolism	1.612	$3.6 \cdot 10^{-3}$	4/5
[c]1460	DNA-directed RNA polymerase, subunit F	J: Translation, ribosomal structure and biogenesis	1.143	$4.3 \cdot 10^{-3}$	4/5
3943	Virulence protein	V: Defense mechanisms	5.564	$4.3 \cdot 10^{-3}$	3/4

[a] Statistical significance of the cluster based on empirical distribution of BLS
[b] m/n: m out of n genes of the cluster are annotated with the consensus category
[c] Also predicted by Wolf *et al.* (2001)

the predictions that are not conserved for significantly long time. In summary, we believe that a relaxed notion of clustering and a rigorous statistical test constitute the optimal strategy for identifying gene clusters that are likely functional.

In this study, we focused on *E. coli*, but the conserved clusters from other bacterial species have actually been computed (available in our website). Thus we expect our results would be a good source of data for further analysis. More generally, it has been shown that genome organization, including operon structure, plays an important role in gene regulation in eukaryotic species (Lawrence, 2002; Sproul *et al.*, 2005; Batada and Hurst, 2007; Ben-Shahar *et al.*, 2007). Thus our computational tools, formulated and developed in a general framework, will facilitate the study of genome evolution in more complex organisms.

ACKNOWLEDGMENTS

We would like to thank Dr. Chengxiang Zhai for many helpful discussions, and Kim *et al.* (2005) for providing us the program *Mcgs* for experimental evaluation. This work was supported in part by the U.S. National Science Foundation under awards FIBR-04-25852.

REFERENCES

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In J. B. Bocca, M. Jarke, and C. Zaniolo, editors, *VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago de Chile, Chile*, pages 487–499. Morgan Kaufmann.

Akerley, B., Rubin, E., Novick, V., Amaya, K., Judson, N., and Mekalanos, J. (2002). A genome-scale analysis for identification of genes required for growth or survival of *Haemophilus influenzae*. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 966–971.

Batada, N. and Hurst, L. (2007). Evolution of chromosome organization driven by selection for reduced gene expression noise. *Nat. Genet.*, **39**, 945–949.

Ben-Shahar, Y., Nannapaneni, K., Casavant, T., Scheetz, T., and Welsh, M. (2007). Eukaryotic operon-like transcription of functionally related genes in *Drosophila*. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 222–227.

Bergeron, A., Corteel, S., and Raffinot, M. (2002). The algorithmic of gene teams. In *WABI '02: Proceedings of the Second International Workshop on Algorithms in Bioinformatics*, pages 464–476, London, UK. Springer-Verlag.

Ciccarelli, F. D., Doerks, T., von Mering, C., Creevey, C. J., Snel, B., and Bork, P. (2006). Toward Automatic Reconstruction of a Highly Resolved Tree of Life. *Science*, **311**(5765), 1283–1287.

Dam, P., Olman, V., Harris, K., Su, Z., and Xu, Y. (2007). Operon prediction using both genome-specific and general genomic information. *Nucleic Acids Res*, **35**, 288–298.

Han, J., Pei, J., Yin, Y., and Mao, R. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Min. Knowl. Discov.*, **8**, 53–87.

He, X. and Goldwasser, M. H. (2005). Identifying conserved gene clusters in the presence of homology families. *J Comput Biol.*, **12**, 638–656.

Hoberman, R., Sankoff, D., and Durand, D. (2005). The statistical analysis of spatially clustered genes under the maximum gap criterion. *J Comput Biol*, **12**, 1083–1102.

Huynen, M. and Bork, P. (1998). Measuring genome evolution. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 5849–5856.

Huynen, M., Snel, B., Lathe, W., and Bork, P. (2000). Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.*, **10**, 1204–1210.

Itoh, T., Takemoto, K., Mori, H., and Gojobori, T. (1999). Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Mol. Biol. Evol.*, **16**, 332–346.

Jacob, F. and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol*, **3**(NIL), 318–356.

Keyamura, K., Fujikawa, N., Ishida, T., Ozaki, S., Su'etsugu, M., Fujimitsu, K., Kagawa, W., Yokoyama, S., Kurumizaka, H., and Katayama, T. (2007). The interaction of *DiaA* and *DnaA* regulates the replication cycle in *E. coli* by directly promoting ATP DnaA-specific initiation complexes. *Genes Dev.*, **21**, 2083–2099.

Kheradpour, P., Stark, A., Roy, S., and Kellis, M. (2007). Reliable prediction of regulator targets using 12 *Drosophila* genomes. *Genome Res.*, **17**(12), 1919–1931.

Kim, S., Choi, J.-H., and Yang, J. (2005). Gene teams with relaxed proximity constraint. In *CSB '05: Proc. 2005 IEEE Computational Systems Bioinformatics Conference (CSB'05)*, pages 44–55, Washington, DC, USA. IEEE Computer Society.

Kolesov, G., Mewes, H., and Frishman, D. (2001). Snapping up functionally related genes based on context information: a colinearity-free approach. *J Mol Biol*, **311**, 639–56.

Lathe, W., Snel, B., and Bork, P. (2000). Gene context conservation of a higher order than operons. *Trends Biochem. Sci.*, **25**, 474–479.

Lawrence, J. (2002). Shared strategies in gene organization among prokaryotes and eukaryotes. *Cell*, **110**, 407–413.

Ling, X., He, X., Xin, D., and Han, J. (2008). Efficiently identifying max-gap clusters in pairwise genome comparison. *J Comput Biol.*, **15**, 593–609.

Luc, N., Risler, J., Bergeron, A., and Raffinot, M. (2003). Gene teams: a new formalization of gene clusters for comparative genomics. *Comput Biol Chem*, **27**, 59–67.

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

Onufryk, C., Crouch, M., Fang, F., and Gross, C. (2005). Characterization of six lipoproteins in the sigmaE regulon. *J. Bacteriol.*, **187**, 4552–4561.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G., and Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A*, **96**, 2896–2901.

Pasek, S., Bergeron, A., Risler, J., Louis, A., Ollivier, E., and Raffinot, M. (2005). Identification of genomic features using microsyntenies of domains: domain teams. *Genome Res*, **15**, 867–74.

Price, M., Huang, K., Arkin, A., and Alm, E. (2005). Operon formation is driven by co-regulation and not by horizontal gene transfer. *Genome Res.*, **15**, 809–819.

Price, M., Arkin, A., and Alm, E. (2006). The life-cycle of operons. *PLoS Genet.*, **2**, e96.

Rogozin, I., Makarova, K., Murvai, J., Czabarka, E., Wolf, Y., Tatusov, R., Szekely, L., and Koonin, E. (2002). Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Res*, **30**, 2212–23.

Rogozin, I., Makarova, K., Wolf, Y., and Koonin, E. (2004). Computational approaches for the analysis of gene neighbourhoods in prokaryotic genomes. *Brief Bioinform*, **5**, 131–149.

Salgado, H., Gama-Castro, S., *et al.* (2006). Regulondb (version 5.0): *Escherichia coli* k-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res.*, **1**, D394–397.

Snel, B., Bork, P., and Huynen, M. (2002). The identification of functional modules from the genomic association of genes. *Proc Natl Acad Sci U S A*, **99**, 5890–5895.

Sproul, D., Gilbert, N., and Bickmore, W. (2005). The role of chromatin structure in regulating the expression of clustered genes. *Nat. Rev. Genet.*, **6**, 775–781.

Tamames, J. (2001). Evolution of gene order conservation in prokaryotes. *Genome Biol.*, **2**, RESEARCH0020.

Tatusov, R., Fedorova, N., Jackson, J., Jacobs, A., Kiryutin, B., Koonin, E., Krylov, D., Mazumder, R., Mekhedov, S., Nikolskaya, A., Rao, B. S., Smirnov, S., Sverdlov, A., Vasudevan, S., Wolf, Y., Yin, J., and Natale, D. (2003). The cog database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.

Watanabe, H., Mori, H., Itoh, T., and Gojobori, T. (1997). Genome plasticity as a paradigm of eubacteria evolution. *J Mol Evol*, **44 Suppl**, S57–64.

Wolf, Y., Rogozin, I., Kondrashov, A., and Koonin, E. (2001). Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res*, **11**, 356–72.

Wu, H., Mao, F., Su, Z., Olman, V., and Xu, Y. (2005). Prediction of functional modules based on gene distributions in microbial genomes. *Genome Inform*, **16**, 247–259.

Yang, Q. and Sze, S.-H. (2008). Large-scale analysis of gene clustering in bacteria. *Genome Res.*, **18**(6), 949–956.

Zheng, Y., Anton, B. P., Roberts, R. J., and Kasif, S. (2005). Phylogenetic detection of conserved gene clusters in microbial genomes. *BMC Bioinformatics*, **6**(NIL), 243.