Abstract. We present a parametrized definition of gene clusters that allows us to control the emphasis placed on conserved order within a cluster. Though motivated by biological rather than mathematical considerations, this parameter turns out to be closely related to the maximum bandwidth parameter of a graph. Our focus will be on how this parameter affects the characteristics of clusters: how numerous they are, how large they are, how rearranged they are and to what extent they are preserved from ancestor to descendant in a phylogenetic tree. We infer the latter property by dynamic programming optimization of the presence of individual edges at the ancestral nodes of the phylogeny. We apply our analysis to a set of genomes drawn from the Yeast Gene Order Browser.

1 Introduction

The definition of synteny blocks, gene clusters or similar constructs from the comparison of two or more genomes entails a trade-off of great consequence: if we place emphasis on identical content and order of the genes, segments or markers in a block or cluster, only relatively small regions of the genome will satisfy this restrictive condition, giving rise to a plethora of tiny blocks while missing large regions common to the genomes. On the other hand, by allowing unrestricted scrambling of genes within blocks (e.g., max-gap \[1\] or “gene teams” \[7\]), we forgo accounting for local genome rearrangement, missing an important aspect of evolutionary history, or we relinquish the possibility of pinpointing extensive local conservation, where this exists.

In this paper, we present a parametrized definition of gene clusters that allows us to control the emphasis placed on conserved order within a cluster. Though motivated by biological rather than mathematical considerations, this parameter turns out to be closely related to the maximum bandwidth parameter of a graph. Our focus will be on how this parameter affects the characteristics of clusters: how numerous they are, how large they are, how rearranged they are and to
what extent they are preserved from ancestor to descendant in a phylogenetic
tree. We infer the latter property by dynamic programming optimization of
the presence of individual edges in a generalized adjacency graph abstractly
representing chromosomal gene order. We apply our analysis to a set of genomes
drawn from the Yeast Gene Order Browser (YGOB) [3]. Among the results, we
find strong evidence for setting a certain fixed value to the cluster parameter.
We also find that we can recover almost all the clusters that can be found
without order constraints, i.e., by the max-gap criterion, indicating that local
order conservation is a lot greater than that unconstrained definition would
suggest.

2 Definitions

Our characterization of gene clusters is made up of a general part that identifies
clusters of vertices common to two graphs, and a specific part where a graph is
determined by the proximity of genes on the chromosomes of a genome. This is
illustrated in Figure [1]

Definition 1. Let $G_S = (V_S, E_S)$ and $G_T = (V_T, E_T)$ be two graphs with a
non-null set of vertices in common $V = V_S \cap V_T$. We say a subset of $C \subseteq V$ is
an $ST$-cluster if it consists of the vertices of a maximal connected subgraph of
$G_{ST} = (V, E_S \cap E_T)$.

Definition 2. For the purposes of genome comparison, we may consider $V_X$ to
be the set of genes in the genome $X$. For genes $g$ and $h$ in $V_X$ on the same
chromosome in $X$, let $gh \in E_X$ if the number of genes intervening between $g$
and $h$ in $X$ is less than $\theta$, where $\theta \geq 1$ is a fixed neighbourhood parameter.

These definitions of edge sets and $ST$-clusters decompose the genes in the two
genomes into identical sets of disjoint 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 genes in this paper. When $\theta = 1$, a
cluster has exactly the same gene content and order (or reversed order) in both
genomes. When $\theta = \infty$, the definition returns simply all the synteny sets, namely
the sets of genes in common between two chromosomes, one in each genome.

Let $\Pi$ be the set of all orderings of $V$. Recall that the bandwidth of a graph
$G(V, E)$ is defined to be

$$B(G) = \min_{p \in \Pi} \max_{u, v \in E} |p(u) - p(v)|.$$  (1)

In a genome $S$ each chromosome $\chi$ determines a physical order among the
genes it contains.

Proposition 1. Bandwidth $B(G_S) = \theta$, as long there are at least $2\theta + 1$ genes
on some chromosome $\chi$ in genome $S$. 