k-mean hierarchical clustering: detailed algorithm and parameter adjustment.

Basic statements

Spacers are defined by their 32-nucleotide sequences. A large number (up to $0.5\mathrm{x}10^{7}$) of spacers needs to be clustered into an initially unknown number of groups, so that spacers in each group are similar to each other and different from spacers from other groups. Also, identical spacers derived from the same protospacer but differing in their orientation (reverse complementary) (Erdmann & Garrett 2012; Lopez-Sanchez *et al.* 2012; Mick *et al.* 2013; Shmakov *et al.* 2014) and spacers produced by imprecise excision (Savitskaya *et al.* 2013), need to be combined and handled together.

A spacer α with a given nucleotide sequence is denoted by the 32 × 4 = 128-dimensional numerical vector Sα, in which information about each nucleotide is stored in 4 corresponding dimensions in the following way:

- base A is denoted as (1, 0, 0, 0).
- base G is denoted as (0, 1, 0, 0).
- base C is denoted as (0, 0, 1, 0).
- base T is denoted as (0, 0, 0, 1).

For example, sequence [AGGC, . .] corresponds to (1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, . .).

While a vector describing a single spacer $S = (S_1, \ldots, S_{128})$ contains only 0s and 1s, the position $C=(C_1,\ldots,C_{128})$ of the center of a cluster, defined as the arithmetic mean of vectors S_α of constituent n spacers,

$$C_j = \frac{1}{n}\sum_{\alpha=1}^{n} S_{j,\alpha},\ j = 1, \ldots, 128, \tag{1}$$

generally is characterized by real numbers $0 \leq C_j \leq 1$.

The distance $D_{\alpha\beta}$ between two spacers or clusters α and β is defined as a sum over 128 dimensions of the absolute value of the difference between their coordinates,

$$D_{\alpha\beta} = \sum_{j=1}^{128} |C_{j,\alpha} - C_{j,\beta}|. \tag{2}$$

This distance is twice the Hamming distance between spacers, since each replacement of a nucleotide removes 1 from the position of the old base and adds 1 to position corresponding to a new base. The radius of a cluster is defined as the distance from its center to its most remote member.

Sorting into tree-like hierarchy

To reduce the amount of data and accelerate the search, we cluster the spacers into a 3-level branching structure with each subsequent level having clusters of progressively higher similarity between members. At the last level of segregation, clusters have radii approximately equal to 3, which reflect the maximum number substitutions corresponding to biologically similar spacers and sets the "resolution limit" of the process. Parameters defining branching were varied and after several experiments we converged to values listed below. The procedure of placing a new spacer into the system of clusters consists of the following steps:

- The first spacer forms the root, the first-level branch, and the second-level branch of the first tree.
- Each new spacer is first matched with the closest tree root. If no tree is found within a distance of 27, the new spacer forms the root, first-, and second-level branches of a new tree.
- If a matching tree is found, the new spacer is then matched with the closest first-level branch coming out from the root. If no first-level branch is found within a distance of 9 from the spacer, the spacer forms new first-level and second-level branches.

- If the matching first-level branch is found, the spacer is then compared to the second-level branches emanating from the first-level branch. It joins the closest second-level branch, and if no such branch exists within a distance of 3 from the spacer, it forms a new second-level branch.

Thus, in such fully developed hierarchy, a spacer is defined by its membership in a tree, in a first-level branch, and in a second-level branch or “final” cluster. The hierarchical scheme allowed us to substantially speed up the search of the target cluster for each new spacer.

This clustering procedure is repeated several times from the beginning, taking into account the results of the previous rounds of clustering. A new round starts with clustering of spacers, which belong to the largest final cluster of the largest branch of the largest tree. Next, spacers from the second largest cluster are re-clustered, etc. After the second iteration the cluster tree does not change significantly. Naturally, some of the clusters may have final radii smaller than the threshold value of 3, while others may contain spacers that are further than 3 substitutions away from the center of their cluster. The latter happens when a spacer, initially within the distance of 3 from the center, becomes further separated as the center moves away due to subsequent addition of new members. We surmise that such “swelling” of clusters has little effect on the final result since if such swollen clusters were broken, most probably, they would have merged during the second stage of clustering.

Shifting, flipping, and merging clusters

The first procedure allows us to reduce the amount of data, which is now represented by sizes and coordinates of centers of a few thousand clusters with radii ≈ 3. Next, we compute pairwise distances between all clusters, taking into account possible reversions (Erdmann & Garrett 2012; Lopez-Sanchez *et al.* 2012; Mick *et al.* 2013; Shmakov *et al.* 2014) and shifts of their sequences. When comparing one cluster to another, we first compute the distance between two sequences in their original form, then for one sequence shifted by ±1 and ±2 bases, and finally we “flip” one sequence, generating a reverse complement sequence and repeat the procedure, looking for the best match. Flips have no distance penalty, but a shift by a single base in either direction adds a 2 to the distance between clusters. In the end, we compute the adjacency matrix of the complete graph where nodes are clusters and edges are labeled by distances between nodes. For a given cutoff distance D, all edges with distances larger than D are removed, normally breaking the complete graph into several disconnected components. Each component is then declared to be a secondary cluster, characterized by its center and the number of constituent spacers. Naturally, the smaller threshold D yields more such secondary clusters; the plot of the number of secondary clusters N vs. D is shown in Fig. S1.

It follows from Fig. S1 that for $5 \leq D \leq 10$, the dependence of N on D is the weakest, which suggests that the natural inter-cluster separation falls into this range. For final clustering of our data, we chose $D = 7$ which is in the middle of this range.

Concluding remarks

Overall, our clustering method offers two main advantages for large CRISPR spacer sets analysis:

- It is significantly faster.

- Compared to clustering based on pairwise BLAST scores, it naturally and simply shows the sequence composition of each cluster and reveals the variability of each nucleotide within the cluster.

References

Erdmann S, Garrett R a. (2012) Selective and hyperactive uptake of foreign DNA by adaptive immune systems of an archaeon via two distinct mechanisms. *Molecular Microbiology*, **85**, 1044–1056.

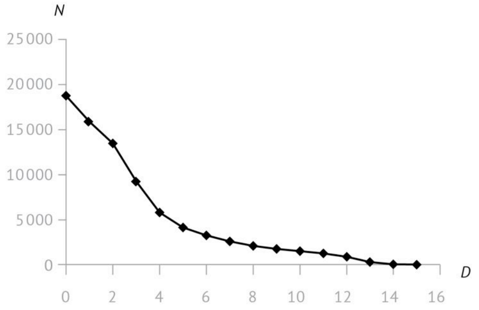
Lopez-Sanchez MJ, Sauvage E, Da Cunha V *et al.* (2012) The highly dynamic CRISPR1 system of Streptococcus agalactiae controls the diversity of its mobilome. *Molecular Microbiology*, **85**, 1057–1071.

Mick E, Stern A, Sorek R (2013) Holding a grudge: persisting anti-phage CRISPR immunity in multiple human gut microbiomes. *RNA Biol*, **10**, 900–906.

Savitskaya E, Semenova E, Dedkov V, Metlitskaya A, Severinov K (2013) High-throughput analysis of type I-E CRISPR/Cas spacer acquisition in E. coli. *RNA biology*, **10**, 716–25.

Shmakov S, Savitskaya E, Semenova E *et al.* (2014) Pervasive generation of oppositely oriented spacers during CRISPR adaptation. *Nucleic Acids Research*, **42**, 5907–5916.

Figure S1. A plot showing the dependence of the number of secondary clusters N vs. the cutoff distance between clusters D.


N
25000
20000
15000
10000
5000
0
D
0
2
4
6
8
10
12
14
16