Model description

Sergio Peignier

November 22, 2016

Abstract

In this report we describe some ideas for a bio-inspired subspace clustering algorithm for data streams

1 Operations in biology

1.1 Horizontal Gene Transfer

- Occurs at very high rate.
- Incorporation in the genome:
  - Comobilization of a genetic region required for common function (not only one gene but functional sets of genes)
  - Can recombine with similar portions in the genome
  - Gene conversion: Gene repairment by homologus recombination of HGT (corepair)
  - Can lead to duplicate: random insertion in genomes and large range of sequence similarity relative to the native homolog.
  - Can be inserted in a different place in the genome thanks to insertion sequences or plasmids.
- Acquisition of novelty:
  - Increase in protein family sizes are more related to HGT (gain of new functions)
  - HGT share less protein-protein interactions and regulatory mechanisms: Very suitable for new clusters
  - Can integrate a new entire function

1.2 Amplification

- Mechanism:
  - Tandem duplication can create side by side several gene copies.
  - Most amplicons are stable and amplify at low rate but amplification generates tandem repeats that recombine at high frequencies.
  - Tandems are very unstable => deletion, amplification, recombination.
  - Recombination is more frequent between close genes and similar genes
  - Rate-limiting step => initial amplification
  - Deamplification often occurs quickly after:
    - removal of selective pressure
    - Acquisition of compensatory mutation
- Compensate genes with weak function:
  - Increase adaptative mutation rate.
  - Gene with weak function of set of genes => co-amplified => restore function.

  - Amplified regions offer a larger mutation target (more mutations accumulating in the amplified set of genes than in a single-gene locus).
  - Recombination between repeats may engage the action of an error-prone polymerase and thus become locally mutagenic.
  - Repeats will accumulate changes and allele reassortment may occur within chromosomes by recombination between the copies, accelerating the combination of favorable changes in one gene.
  - Acquisition of compensatory mutation
- Rudimentary form of regulation:
  - Reversible change in dosage
  - Enable adaptation to variable environments
- Acquisition of novelty:
  - Duplication and divergence (subfunctionalization, neofunctionalization)

1.3 Horizontal gene transfer and amplification

- HGT genes are typically maladapted for expression on the host => Through amplification => adapt more rapidly by the joint action of three effects.
- HGT usually undergo Amplification and usually lead to paralog retention
- A gene inserted through HGT next an existing gene is ready for amplification.
- New functions acquired may undergo amplification.

1.4 Duplicates

- Repeats and instability:
  - Multiple repeats increase homologus recombination and thus increase instability:
    - variation of number of duplicates generated
    - mutations
  - Repeats undergo conversion which tend to keep them alike, which lead to more recombination and thus increases instability through recombination.
  - When frequent stress => generating variability at reduced number of loci become beneficial => This is achieved by maintaining repeats.
- Divergence of duplicates.
  - Reduce instability:
    - Reduce homologus recombination of repeats
    - Reduce gene conversion
  - Selection of duplicate genes (paralogs) that diverged and obtained a new function:
    - subfunctionalization: The two copies specialize in subfunctions of the primitive gene Alows to escape from adaptative conflict (2 functions cannot be optimized on the same gene). By splitting the two functions => independence of evolution.
    - neofunctionalization: One copy acquire a new function. Close related to the function required. Novelty built on a conserved enzymatuc mechanism If selection for amplification is stable: Genes have more time to diverge and adress a new function (related to the first one) and be selected. This new function may be inefficient and need
- Silencing
  - Deletion: Arise from recombination between repeats at high rates
  - Silencing by mutational inactivation: Accumulation of point substitution => pseudogene (too diferent to act as a repeat and to recombine)

1.5 Genome size

- Number and density of IS increases rapidly with genome size (so bigger genome implies more recombination and mutations)
- Amplicons are stable until they are turned into tandem (limiting step)
- Only when turned into tandems they become unstable
- Deletion => back again to limiting step
- Most genes may have arise from HGT and large genomes tend to contain more distant phylogenetically discordant genes.
  - Illegitimate recombination rate per gene increases in large genomes since such genome contain more transposons, integrases and phage elements that may facilitate the integration of foreign DNA.
  - Large genome are also composed of plasmids or megaplasmids and thus have higher rates of translocation.
  - Modularity may also increase the size and efficiency of HGT
  - Genome size is correlated to biodiversity: More complex ecological interactions in species-rich communities could increase the demand for larger gene repertoires, which are expanded by accepting genes from those phylogenetically distant organisms in the environment.

1.6 Mutators

Deletion of reparing genes => Increase mutation rate in stress times

2 Model description

2.1 Genome structure

Genes description Each gene can be perceived as an abstract sequence that codes for a molecular product and that the combination of molecular products associated together codes for a function. In this case a function can correspond to the location of a core-point along a given dimension or to the location of an entire core-point in its own subspace. The genome considered here is a coarse-grained genome and the underlying gene sequence is never represented explicitly.

Co-localization of genes As reported for instance in [RN04], bacterial genes sharing a common function or encoding interacting proteins are often co-localized in the genome (e.g., operons and gene clusters) and such an organization appears to be conserved in different organisms despite the frequent shuffling and rearrangement operations genomes undergo. In [RN04], authors suggest that such phenomenon provide a selective advantage specially under variable environmental, since this phenomenon should facilitate the acquisition of entire new traits (through Horizontal Gene Transfer) and the co-amplification of genes involved in the same function (through Tandem Duplication).

In order to allow for HGT and co-amplification of functionally related genes we will consider that all genes involved in the same function are assumed to be co-localized in the genome: Genes encoding the location of the coordinate of a core-point along a given dimension are all contiguous and so are genes contributing to the location of a given core-point. However the order of the genes encoding the same function and the order between groups of genes encoding the same function are not represented in the model **REM** **On a besoin d'une description mathematique:** https://en.wikipedia.org/wiki/Family_of_sets.

Representation of repeats Genes that are repeated in tandem arrays, tend to undergo conversion which tends to keep them alike, therefore we are going to consider that the $\mathcal{W}_{c,d}$ genes involved in describing the location of the core-point c along dimension d, are indistinguishable. With this assumption, the model does not require to save the value of each gene. It is sufficient to save the number of genes involved in a particular trait (core-point coordinate) and the trait itself.

Representation of non-functional genes Genomes in biology are not only composed by functional genes, they also contain sequences of non-coding DNA that do not encode any protein. The genome sizes and the proportion of non-coding DNA can vary greatly between species. In this model we wont encode the content each non-functional gene, we are only going to consider the number of non-functional genes, assuming that they can be distributed uniformly between the clusters of genes that encode each function (core-point). According to the literature the number of elements facilitating the rate if illegitimate recombination and integration for foreign DNA (e.g., transposons, integrases, phage elements and plasmids) increases with the genome size. Therefore, having a non-functional may allow individuals to tune their mutation rate, without the side effects that could have an increase of the number of functional genes.

2.2 Horizontal gene transfer (HGT)

According to [Law03, RBG+96, LR96], co-localization of functionally related genes could enhance the adaptation of organisms to new environments through the acquisition of entire new traits through Horizontal Gene Transfer (HGT), since nearby genes are more likely to be transferred together.

2.2.1 Number of HGT per replication

According to the literature, the number of elements facilitating the integration of foreign DNA increases with the genome size. Let μ_{int} denote the probability that a gene contains such elements in its immediate neighborhood. Then assuming that each one of these elements is independent from each other, we can assume that the number of HGT events per replication will follow a binomial distribution $\mathcal{B}(\mu_{ins}, W_f + W_{nf})$ where $W_f = \sum_{c,d} \mathcal{W}_{c,d}$ and W_{nf} denotes the number of non-functional genes.

2.2.2 Size of the HGT

Given that an entire genetic region is co-mobilized we could insert several dimensions at once. In order to define the size of the HGT, let us assume that each gene has a probability μ_{hgt} to be transmitted. If we assume that each gene can be transmitted independently from others, then the number of genes transmitted follows a Binomial distribution $\mathcal{B}(D, \mu_{hgt})$ where D is the dimensionality of the (data) point originating the HGT event.

2.2.3 Point of insertion of the HGT

An HGT can be used to modify the location of an existing core-point or can be used to build a new one. The insertion of the HGT in the genome could be explained in different ways:

Insertion by similarity and distance between HGT and core-points We know that HGT can be inserted in the genome according to similarity (conversion or insertion). Therefore we could measure some distance between the HGT and the core-point in the subspace of the HGT, and then compute the probability to assign the HGT to the given core-point as a function of its distance. The distance to zero could be used to compute the probability to build a new core-point using the HGT (this option has not been detailed here).

HGT as functional set of genes and association to function Each HGT can be perceived as a set of genes that are functionally related (approximation of the coordinates of a core-point in a given subspace). The function tackled by the genes acquired through HGT may either already exist in the genome or be a new one. Thus we should either associate the HGT genes to the appropriate core-point (first case) or create a new core-point that captures the new function (second case). Whenever a random set of genes is received, it is either associated to an existing core-point or it is chosen to define a new core-point. The probability to associate the HGT to an existing core-point depends both on its distance to the core-point size and on the size of the core-point.

Definition of the probability of insertion Intuitively a HGT extracted from a point that is far away from a core-point is unlikely to be helpful to improve the core-point location. A HGT extracted from a point that is far away from all the existing core-points is more likely to be useful a new core-point (or it is a noisy point).

Moreover a core-point with a high weight is likely to have a high contribution to the gain in SAE, and is also likely to contain coordinates with higher variances and more-points. Therefore it seems reasonable to attribute more exploration to such core-points.

Let x denote the point from which the HGT is extracted, let $\mathcal{L}_c$ denote the location of the core-point c and $\mathcal{W}_c = \sum_d \mathcal{W}_{c,d}$ its weight (number of genes associated to it) and let $W = \sum_{c,d} \mathcal{W}_{c,d}$.

- Let $q_c(x) = exp(-\frac{dist(\mathcal{L}_c,x)}{\mathcal{W}_c})$
- Let $q(x) = exp(-\frac{W}{dist(0,x)})$
- Let $Q(x) = q(x) + \sum_c q_c(x)$
- Let $p_c(x) = \frac{q_c(x)}{Q(x)}$ be the probability that the HGT is associated to the core-point c and let $p(x) = \frac{q(x)}{Q(x)}$ be the probability that the HGT is used to define a new core-point.

Intuitively the higher the distance to a given core-point and the lower the weight of the core-point, the lower the probability to associate the HGT to it. Conversely the lower the model size and the higher the distance to zero, the higher the probability to define a new core-point.

2.2.4 Analysis of the probability function

In order to analyze the impact of different variables on the probability function, we simulate The existence of C core-points chose uniformly between -1 and 1 in a D dimensional space and having subspaces with a relative size of d percent of the complete space. Let us consider that each core-point coordinate is encoded by the same number of genes $W_{c,d} = w$ for all c and d in the model. We also generate a set of points in order to simulate a dataset, the data points are generated either uniformly or following a gaussian distribution around a chosen core-point with a standard deviation equal to 0.1.

The distance between each core-point and the data points is computed using the square of the euclidean distance, in order to be consistent with the search of centroids.

We modified one by one the major variables of the system, the number of core-points C, the dimensionality of the space D and the number of genes w used to encode each coordinate. Notice that w and d had a very similar effect, and thus only the results obtained on w are showed here. The default values used in each case are $C = 5$, $d = 0.1$, $D = 30$ and $w = 1$.

Impact of the number of core-points When the point follows a gaussian distribution around a core-point, and thus is likely to be close to it, it is more likely to associate the point to its corresponding core-point. When the number of core-points increase, the probability to associate the point to the correct core-point decreases, since other core-points could appear to be close enough and then lead to a decrease in the probability.

When we deal with uniformly distributed points, the farthest the data point, the higher the probability to define a new core-point. However, when there are already many clusters, this probability decreases rapidly and it is more likely to associate the point to another existing core-point. It is specially easy to create new core-points when there are few of them.

Given these probabilities, the number of core-points is likely to converge to a finite value.

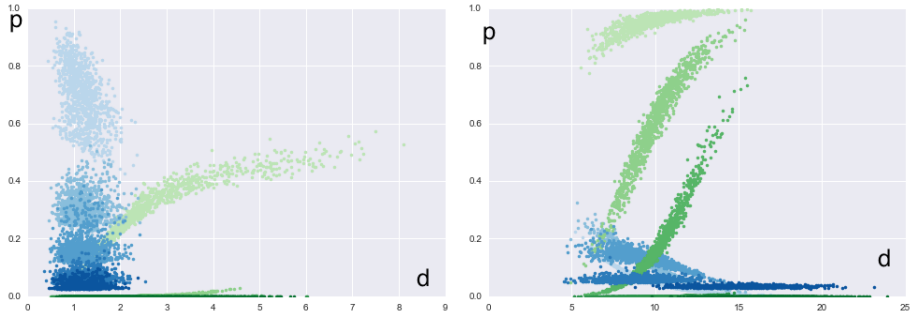

Figure 1: Probability to associate the point (that follows a gaussian distribution around a core-point and a uniform distribution respectively) to its closest core-point (blue) or to define a new core-point (green) as a function of the distance between to point and the given core-point or zero respectively. Results for $C = 1$ (lighter colors), $C = 5$, $C = 10$, $C = 30$ and $C = 50$

Impact of the space dimensionality When the point follows a gaussian distribution around a core-point, and thus is likely to be close to it, it is more likely to associate the point to its corresponding core-point, regardless of the dimensionality of the dataset.

When we deal with uniformly distributed points, the farthest the data point, the higher the probability to define a new core-point. However the dimensionality does not seem to affect the probabilities, since the shape of the curves adapts to the increase of distances related to the increase in dimensionalities such that the probabilities should behave similarly.

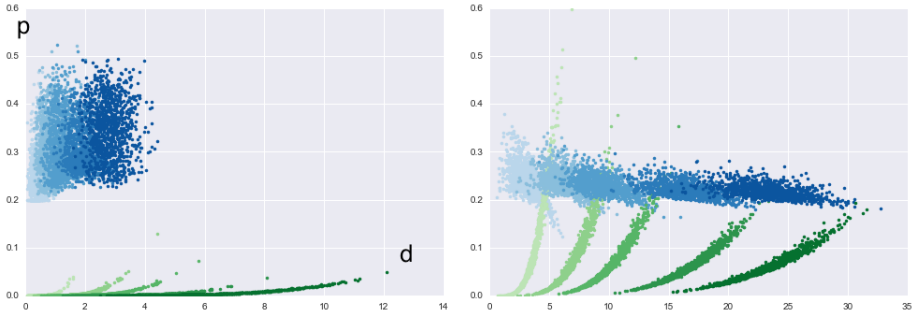

Figure 2: Probability to associate the point (that follows a gaussian distribution around a core-point and a uniform distribution respectively) to its closest core-point (blue) or to define a new core-point (green) as a function of the distance between to point and the given core-point or zero respectively. Results for $D = 10$ (lighter colors), $D = 20$, $D = 30$, $D = 50$ and $D = 70$

Impact of the number of genes per core-point coordinate When the point follows a gaussian distribution around a core-point, and thus is likely to be close to it, it is more likely to associate the point to its corresponding core-point, regardless of the number of genes per core-point coordinate. Notice however that the results obtained for lower values of w are more variable than those obtained for higher values of w.

When we deal with uniformly distributed points, the farthest the data point, the higher the probability to define a new core-point. However the probability to create a new core-point is more reactive for lower values of w since the model weight increases proportional with w in this example, this behavior is interesting since core-points with higher weights are likely to define clusters with more points and more variables and are then more likely to have more distant points.

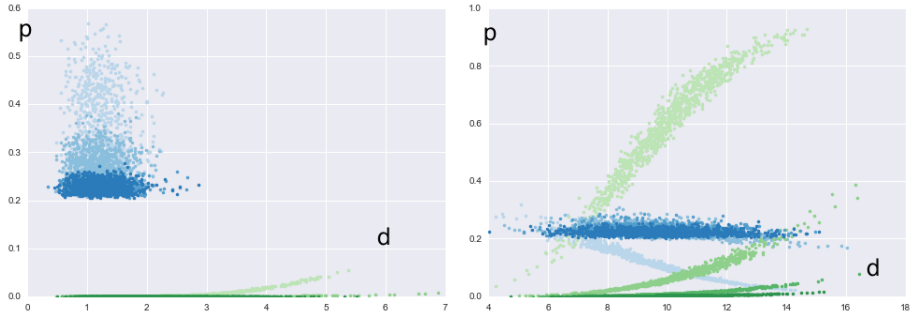

Figure 3: Probability to associate the point (that follows a gaussian distribution around a core-point and a uniform distribution respectively) to its closest core-point (blue) or to define a new core-point (green) as a function of the distance between to point and the given core-point or zero respectively. Results for $w = 1$ (lighter colors), $w = 2$, $w = 3$ and $w = 4$.

2.2.5 Model update

Model update Let c denote the identifier of the core-point to which the HGT has been assigned (or a new one). Then $\forall x_d \in x$:

- $\mathcal{L}_{c,d} \leftarrow \frac{x_d + \mathcal{W}_{c,d}\mathcal{L}_{c,d}}{\mathcal{W}_{c,d}+1}$
- $\mathcal{W}_{c,d} \leftarrow \mathcal{W}_{c,d} + 1$

HGT and Amplification Given the HGT are usually ready to undergo amplification (and usually need it) can we directly start amplification process after insertion?

2.3 Amplification and duplication

[RN04] suggests that nearby genes are also more likely to undergo co-amplification, this should result in a rudimentary form of regulation leading to dosage-effect advantage. Moreover [CW08] reports another beneficial potential role of amplification. According to the authors genes may encode minor activities besides their main function (promiscuous protein activities). Such minor activities are likely to be inefficient, however amplification could cope with inefficiency through dosage effects. Finally since many copies are required to tackle the required function, there are higher chances of acquisition of compensatory mutations that could lead to more efficient activity.

2.3.1 Probability of amplification

An initial duplication (unequal sister strand exchange, polymerase slippage) generates two adjacent identical copies of a region (the amplicon) that can undergo amplification (through homologous (RecA-dependent) recombination). Since single genes are stable, the initial duplication is the limiting-step in amplification. Once duplicated, the amplicons undergo more easily Tandem duplication which generates contiguous arrays of duplicated genes. These regions are unstable and are likely to be amplified, deleted or mutated.

Amplification occurs between two repeated regions. If we assume that each pair of repeats has a probability μ_a to lead to an amplification, and that the probability of each pair is independent from the other ones, then the probability to undergo amplification follows a binomial distribution $\mathcal{B}(C, \mu_a)$ where $C = \binom{\mathcal{W}_{c,d}}{2} = \frac{\mathcal{W}_{c,d} \times (\mathcal{W}_{c,d}-1)}{2}$ denotes the number of possible combinations of the $\mathcal{W}_{c,d}$ repeats that define the location of core-point c along dimension d. Notice that only regions that contain at least 2 repeats are able to undergo amplification. The total number of amplifications follows the binomial distribution $\mathcal{B}(\sum_{c,d} \frac{\mathcal{W}_{c,d} \times (\mathcal{W}_{c,d}-1)}{2}, \mu_a)$

2.4 Probability of duplication

An initial duplication is necessary to generate two adjacent identical copies of a region (the amplicon) that can then undergo amplification. Let us assume that each gene has the same probability μ_{dup} to undergo a single gene duplication. Assuming that each gene is independent of each other,

the number of duplications follows the binomial distribution $\mathcal{B}(\sum_{c,d} \mathcal{W}_{c,d}, \mu_{dup})$. Non-functional genes are also able to undergo duplication, in this case the number of non-functional elements duplicated follows a binomial distribution $\mathcal{B}(\sum_{c,d} W_{nf}, \mu_{dup})$, and the amount of non-functional elements is updated as follows: $W_{nf} \leftarrow W_{nf} + \mathcal{B}(\sum_{c,d} W_{nf}, \mu_{dup})$

Notice that a gene acquired by HGT that is added to an existing function facilitates the appearance of a subsequent amplification without needing to wait for a single duplication event.

2.4.1 Size of amplification

Since Genes encoding the same core-point coordinate are considered to be contiguous in the genome, the amplification operation may be modeled as the duplication of the genome fragment defined within two break-points chosen uniformly between any of the repeated genes. Moreover since repeated genes are considered as identical, this operation is equivalent to a multinomial draw with 3 classes (the fragment at the left of the duplication fragment, the duplication fragment and the fragment at its right). Since by definition, genes outside the duplication fragment are not duplicated, the number of duplicated genes follow simply a binomial distribution $\mathcal{B}(1/3, n_g)$, where $1/3$ reflects the probability to fall in the duplication fragment (rather the gene is in one of the two regions outside the duplication fragment or it is in the duplication fragment) and n_g denotes the number of genes in the amplicon (e.g., $\mathcal{W}_{c,d}$ for the genes denoting the coordinates of center c along dimension d).

2.4.2 Model update

Let n denote the number of genes that are duplicated during an amplification (or duplication) event. Let c be the identifier of the core-point described by these genes and let d be the dimension along with the core-point is described.

- If the core-point denotes a cluster, then uniformly pick n coordinates $\{x_d^1, \ldots, x_d^n\}$ of points belonging to the cluster associated to the core-point c along the dimension d, otherwise thake n random coordinates of data points along dimension d.
- $\mathcal{L}_{c,d} \leftarrow \frac{\sum_i x_d^i + \mathcal{W}_{c,d}\mathcal{L}_{c,d}}{\mathcal{W}_{c,d}+n}$
- $\mathcal{W}_{c,d} \leftarrow \mathcal{W}_{c,d} + n$

2.5 Deamplification

2.5.1 Probability of deamplification

Since deamplification occurs between two repeated regions, just as amplication, the probability of deamplification is define likewise the one for amplification:

Let us assume that each pair of repeats has a probability μ_{da} to lead to a deamplification, and that the probability of each pair is independent from the other ones, then the probability to undergo deamplification follows a binomial distribution $\mathcal{B}(C, \mu_{da})$ where $C = \binom{\mathcal{W}_{c,d}}{2} = \frac{\mathcal{W}_{c,d} \times (\mathcal{W}_{c,d}-1)}{2}$ denotes the number of possible combinations of the $\mathcal{W}_{c,d}$ repeats that define the location of core-point c along dimension d. Notice that only regions that contain at least 2 repeats are able to undergo deamplification. The total number of deamplifications follows the binomial distribution $\mathcal{B}(\sum_{c,d} \frac{\mathcal{W}_{c,d} \times (\mathcal{W}_{c,d}-1)}{2}, \mu_a)$

2.6 Probability of deletion

Let us assume that each gene has the same probability μ_{del} to undergo a single gene deletion. Assuming that each gene is independent of each other, the number of deletions follows the binomial distribution $\mathcal{B}(\sum_{c,d} \mathcal{W}_{c,d}, \mu_{del})$. Non-functional genes are also able to undergo deletions, in this case the number of non-functional elements deleted follows a binomial distribution $\mathcal{B}(\sum_{c,d} W_{nf}, \mu_{del})$, and the amount of non-functional elements is updated as follows: $W_{nf} \leftarrow W_{nf} - \mathcal{B}(\sum_{c,d} W_{nf}, \mu_{del})$

2.6.1 Size of deamplification

Since Genes encoding the same core-point coordinate are considered to be contiguous in the genome, the deamplification operation may be modeled as the deletion of the genome fragment defined within two break-points chosen uniformly between any of the repeated genes. Moreover since repeated genes are considered as identical, this operation is equivalent to a multinomial draw with 3 classes (the fragment at the left of the duplication fragment, the duplication fragment and the fragment at its right). Since by definition, genes outside the duplication fragment are not duplicated, the number of duplicated genes follow simply a binomial distribution $\mathcal{B}(1/3, n_g)$, where $1/3$ reflects the probability to fall in the deletion fragment (rather the gene is in one of the two regions outside the deletion fragment or it is in the deletion fragment) and n_g denotes the number of genes in the amplicon (e.g., $\mathcal{W}_{c,d}$ for the genes denoting the coordinates of center c along dimension d).

2.6.2 Model update

According to the literature, deletion usually follows compensatory mutation or no more selective pressure. Assuming that a compensatory mutation arises very quickly, we consider that genes can be deleted without leading to a modification of the description of the model at the phenotype level (without modifications of the coordinates matrix) as long as the trait is not fully deleted, i.e., if at least one copy is kept, and then only the weights matrix should be updated.

Let n denote the number of genes to be deleted. Let c be the identifier of the core-point described by these genes and let d be the dimension along with the core-point is described.

- If $\mathcal{W}_{c,d} - n = 0$,:
  - $\mathcal{L}_{c,d} \leftarrow 0$
  - $\mathcal{W}_{c,d} \leftarrow 0$
- Otherwise:
  - $\mathcal{W}_{c,d} \leftarrow \mathcal{W}_{c,d} - n$

2.7 Divergence of duplicates

2.7.1 Single gene mutations

Genes may undergo single point mutation with a probability μ_s. Let us consider the case where a gene contributing to the location of the core-point c along dimension d is mutated. Three outcomes are possible:

Modification of the enzymatic activity In the first scenario, the mutation leads to a modification of the contribution of the gene to the phenotypic function (e.g., the activity of the enzyme encoded in the gene is modified). In this case, the weight matrix is kept unchanged and the matrix describing the location of the core-points is updates in order to take into account the mutation. Let x_d denote a random coordinate picked from a point from cluster of core-point c if it is not empty or a from a random data point otherwise. The coordinate of core-point c along dimension d is modified as follows: $\mathcal{L}_{c,d} \leftarrow \frac{\mathcal{L}_{c,d} \times (\mathcal{W}_{c,d} - 1) + x_d}{\mathcal{W}_{c,d}}$

Neo-functionalization In the second scenario, the mutation leads to the acquisition of a new function (closely related to the initial one). This phenomenon is known as neo-functionalization. In this model we consider that such a mutation leads to an increase of the size of the core-point subspace through the acquisition of a coordinate along a new dimension. The new coordinate $x_{d'}$ is acquired by drawing a random data object that belongs to the core-point cluster (or a random object if the cluster is empty), then a dimension d' of the data point space that does not belong to the core-point subspace is randomly picked. The model is modified as follows:

- $\mathcal{W}_{c,d} \leftarrow \mathcal{W}_{c,d} - 1$
- $\mathcal{W}_{c,d'} \leftarrow \mathcal{W}_{c,d'} + 1$
- $\mathcal{L}_{c,d'} \leftarrow \frac{\mathcal{L}_{c,d'} \times \mathcal{W}_{c,d'} + x_{d'}}{\mathcal{W}_{c,d'} + 1}$

Silencing In the last scenario, the mutation turns the gene into non-functional. This phenomenon is known as gene silencing. According to the literature, silencing usually follows compensatory mutation or lack of selective pressure. Assuming that a compensatory mutation arises very quickly, we consider that genes can be silenced without leading to a modification of the description of the model at the phenotype level (without modifications of the coordinates matrix) as long as the trait is not fully silenced, i.e., if at least one copy is kept, and then only the weights matrix should be updated. Notice that in this case the number of non-functional elements increases by one.

- If $\mathcal{W}_{c,d} - 1 = 0$,:
  - $\mathcal{L}_{c,d} \leftarrow 0$
  - $\mathcal{W}_{c,d} \leftarrow 0$
- Otherwise:
  - $\mathcal{W}_{c,d} \leftarrow \mathcal{W}_{c,d} - 1$
- $W_{nf} \leftarrow W_{nf} + 1$

2.8 Population of subpopulations

2.8.1 Two stage population model

In order to keep a good ratio of exploration and exploitation, we decided to apply a two stage population model, A population is constituted by a fixed number of subpopulations, and each subpopulation is constituted by the progeny of the best individual of the subpopulation of the previous generation. At each generation the different subpopulations are ranked according to the fitness of their respective best individual. The number of children per subpopulation is computed according to an exponential rank selection schema (each subpopulation has at least one children). Then the best individual of each subpopulation reproduces and then the best individual of each subpopulation is kept. If we imagine a selection pressure of 0.5 and 100 subpopulations, then the best individual would have around 50 children, the second best 25, the third 12.5, the fourth 6.25, the fifth 3.125, the sixth 1.5625, the seventh 0.78125 and so on.

2.8.2 Single individual based population

The problem related to such a population of subopulations or even a population with individuals that may come from different lineages, is that recording microclusters with the same identifiers requires more operations that just recording the actual cluster. This problem does not appear for populations generated by the best individual at each generation. If we decide to consider a population based on a single individual, we can support our model using https://en.wikipedia.org/wiki/Evolution_strategy

2.9 General ideas

- Single duplications events (from W=1 to further), creation of new dimensions (neo/sub-functionalization) and HGT lead to amplification.
- Amplification can also arise in tandem genes
- Deamplification has no impact on the location
- Pseudogeneation has no impact on the location
- Number of mutational events is a function of the the genome size
- Using medians will be probably more complicate for aggregations.
- We can have an implicit population as in Kymero
- Given that searching centroids is coherent with minimization of euclidian distance, we will probably need to move towards euclidian distance.

- If we use centroids can we have a fast way to see if the modification is positive or not without computing all distances?
- We can use the variance as in PROCLUS to exclude dimensions that are too spread.
- We could image introducing large duplications and deletions in order to increase the importance of an entire core-point in one operation or to lose a useless cluster in one operation.
- Another nice idea was to keep points in clusters (as DBSCAN core-points) to generate various shapes => more complicate given our little amount of time

References

[CW08] Gavin C Conant and Kenneth H Wolfe. Turning a hobby into a job: how duplicated genes find new functions. *Nature Reviews Genetics*, 9(12):938–950, 2008.

[Law03] Jeffrey G Lawrence. Gene organization: selection, selfishness, and serendipity. *Annual Reviews in Microbiology*, 57(1):419–440, 2003.

[LR96] Jeffrey G Lawrence and John R Roth. Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics*, 143(4):1843–1860, 1996.

[RBG+96] JRea Roth, NICHOLAS Benson, TIMOTHY Galitski, KENNETH Haack, JEFFREY G Lawrence, and LYNN Miesel. Rearrangements of the bacterial chromosome: formation and applications. *Escherichia coli and Salmonella: cellular and molecular biology*, 2:2256–2276, 1996.

[RN04] Andrew B Reams and Ellen L Neidle. Selection for gene clustering by tandem duplication. *Annu. Rev. Microbiol.*, 58:119–142, 2004.