# Abstract

# This is a research project. In order to have a better grasp of the issue that this project is facing and better understand the suggested solution we will start with presenting some background:

# http://upload.wikimedia.org/wikipedia/commons/8/81/ADN_animation.gifDNA: The DNA is a molecule that exists in each and every of the cells in our body, and contains most of the [genetic](https://en.wikipedia.org/wiki/Genetics) instructions used in the development and functioning of all known living [organisms](https://en.wikipedia.org/wiki/Organism), from the simplest bacteria to human beings.

# The DNA consists of two strands coiled around each other. Each of the strands is composed from a sequence of Nitrogenous bases (labeled as A, T, G & C).

# The two strands are connected to each other in a common manner: "A" and "T" will always be connected, and "C" and "G" will always be connected. This feature allows us to reconstruct a strand, given the other strand it was connected to. Every connected pare is referred to as a "base pair"

The surprising fact about DNA is that 99.9% human DNA is common to all human beings. And this is a feature this project heavily relies on.

Figure 1 - The DNA consists of two strands coiled around each other

The data encapsulated within the genome can be represented as string of the bases it is composed of, and in this project we will treat it as such (e.g. TGACCGTCAG...).

The DNA is composed of – digitally wise it is equivalent to a little more the 1.6 GB (or about to CDs…).

Another worth mentioning "feature" of the DNA is that it could mutate. The vast majority of mutations are harmless but is the mutation is located in certain location in the DNA it might cause genetic diseases, and cancer in particular.

These days, when a patient is getting diagnosed for cancer the review process is long and cumbersome:

Figure 2 – A connects to T and G connects to C

1. Sampling the patient DNA.
2. Comparing the samples to a DNA from a healthy source.
3. Locating mutations.
4. Manually searching in medicine databases for information about the mutations and their correlation with cancer.
5. Manually searching for an existing drug that is affective with the specific diagnosed type of cancer.

Among other challenges in the above process the comparing (#2) stands out as a time consuming step. In order to accomplish this task at a feasible time line, hospitals rent server farms for doing the comparisons. And still, this process takes about a full day.

This project will focus on the optimization of this step.

For the optimization we have tested and benchmarked few items:

1. The impact of length of samples on calculation runs time.
2. The impact of number of samples on calculation runs time.
3. The impact of sorting the samples on calculation runs time.
4. The impact of making the process parallel on multiple cores.

After testing and benchmarking, it seems we get best results by making the process parallel in a manner that every sampled DNA sequence is being aligned using a separate thread (i.e. rather than having the alignment algorithm aligned internally).

This type of optimization stays stable with no dependency on samples length/ number / sort.