This project has received funding from the Federal Ministry of Education and Research and the European Union’s Seventh Framework Programme for research, technological development and demonstration under grant agreement no 605728

**Algorithms and data structures on faulty sequences**

**Motivation:**
Errors occur due to noise, corruption of data, or mishaps
- electronically transmitted information (bit stream)
- natural language (spoken or written text in some alphabet)
- in genetics (heredity transmission of DNA, mutated binding of nucleotides, protein synthesis of amino acids)

**Problem:**
Given a big piece of text and some pattern, in the form of sequences of characters, identify efficiently within the given text all (approximate) occurrences of the pattern.

**Directions:**
- combinatorics on words
- data structures
- algorithms

**Data structures:**
- (new) suffix arrays
- (new) suffix trees
- k-encodings
- longest previous factor
- longest common prefix
- Rauzy graphs

**Algorithms:**
- classic pattern matching
- jumbled pattern matching
- histograms analysis

**Combinatorics on words:**
- weak and strong periodicity of words
- non-transitivity of the compatibility relation
- approximations of the “three-squares lemma”
- bounds on the number of repetitions

**Applications:**
- enrichment of the research area and connections with related fields
- improvement of already existing approximation algorithms regarding biological sequences alignment (global or local)
- better natural language processing tools
- improvement of the DNA sequencing algorithms by improving the time of the assembly of reads

**Objectives:**
- the investigation of the pattern matching problem in the setting of faulty sequences with bounded errors having as focus the improvement of the time and/or space complexity.
- the investigation of the indexing problem in the presence of bounded errors with focus on the construction of data structures that reduce the preprocessing and/or query time.
- the investigation of the k-equivalence pattern matching problem in a dynamic settings with focus on the reduction of the running time of the online algorithm.

**Solutions:**
- avoid them
- assimilate and recognize them