Genetic Heterogeneity

Version : 0.2
Original Date : May 8, 2013
Revision Date : May 8, 2013

Software Requirements Specification

Authors :

David Lauzon
Alain April

Département de génie logiciel et des TI

Université du Québec
École de technologie supérieure

Client :

Dr. Mark Trifiro

Endocrinology Department

Jewish General Hospital

Revision History

Date	Version	Description	Author
8 mai 2013	0.2	Sections introduction, use cases (UC-01 "Locate an enzyme in a DNA sequence", and started "UC-02 Compare two DNA sequences"), and actors.	David Lauzon
5 mai 2013	0.1	Gabarit initial	David Lauzon

Table of Contents

List of Figures

1 Introduction

1.1 Objective

The purpose of this SRS is to fully describe the external behavior of the software application for the Genetic Heterogeneity project, as well as nonfunctional requirements, design constraints, and other factors necessary to provide a complete, comprehensive description of the software requirements.

1.2 Scope

(TODO: in progress)

1.3 References

This sub-section list all project-related references or applicable documents that bear on this project.

- (more to come...)

1.4 Hypothesis and Dependencies

An assumption is made that the DNA sequences provided to the system, are already sequenced using an external software.

Also, the file format of the DNA sequences is provided as a simplified FASTA format, e.g. a FASTA format[1] with a single DNA sequence and without a description (otherwise, please precise that this is required).

Other than the sequencer, this application does not require other external software.

1 http://en.wikipedia.org/wiki/FASTA_format

2 Overview of Use Cases

This section provides an overview of the use-case model.

2.1 High-level diagram of all use cases

Figure 2.1 is a diagram that shows a high-level overview of the entire use-case model, and more precisely it shows by which actor each use case will be used.

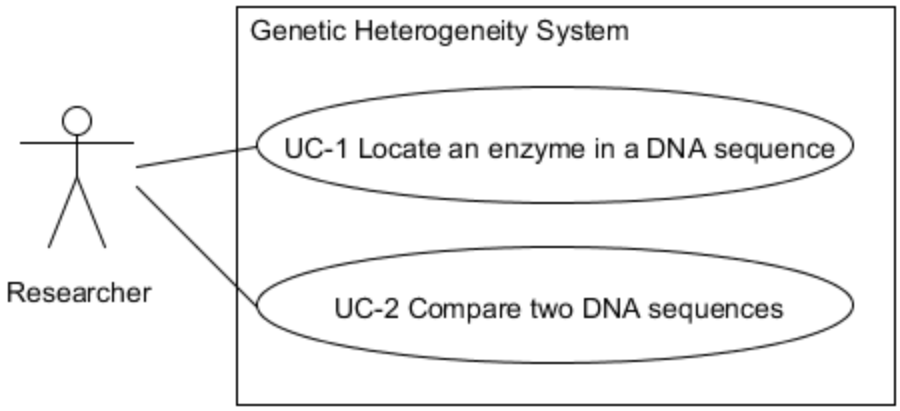

Figure 2.1: *Use Cases Overview Diagram*

2.2 Description of the use cases

This section lists the major use cases that can be fulfilled by using the system:

UC-1 Locate an enzyme in a DNA sequence

With an enzyme pattern, given as an input to the system, the user wants to locate all occurrences of the enzyme pattern within the source DNA sequence (whole genome, chromosome, or other DNA sub-sequence).

An enzyme pattern is the sequence of nucleotides that the enzyme is looking for (base pattern), and may also includes additional preceding or following (unknown) nucleotides. Note, that the cell can scan the DNA sequence in both directions, and in that case the whole pattern including the "unknown" nucleotides is reversed. Once the pattern is found, both the base pattern and the additional nucleotides are "cut" (e.g. duplicated for cellular activity).

For each occurrence of an enzyme pattern, the system must compute the distance in number of nucleotides to the next occurrence.

Figure 2.2: *Bbs-I Enzyme Pattern (Source: NEB.com)* [2]

For example, Figure 2.2 above shows the pattern for the Bbs-I enzyme. When scanning in the *left-to-right* direction, the cut nucleotides are:

- GAAGAC, and the following 2 nucleotides to the *right* on the 5'-3' strand.
- CTTCTG, and the following 6 nucleotides to the *right* on the 3'-5' strand.

When scanning in the *right-to-left* direction, the cut nucleotides are:

- GTCTTC, and the following 6 nucleotides to the *left* on the 3'-5' strand.
- CAGAAG, and the following 2 nucleotides to the *left* on the 5'-3' strand.

Figure 2.3 below gives a complete example of the DNA sections that would be cut by the Bbs-I enzyme.

2 https://www.neb.com/products/r0539-bbsi

Figure 2.3: *DNA Cutting Example for Bbs-I enzyme.*
The search pattern is in yellow, and the DNA that is cut is both the green and yellow.
Note: the DNA sample is taken from real data, except that the distance between the occurrences is about 20 times larger in the real data than in this example above.

The pattern occurrences cut the DNA sequence into fragments. Note that the largest boundaries of the enzyme pattern on either strand are used for this operation.

Figure 2.4 below shows the fragments resulting from cutting the example DNA sequence of Figure with the Bbs-I enzyme.

#	Ends	Coordinates	Length (bp)
1	BbsI-BbsI	258-399	142
2	BbsI-BbsI	400-535	136
3	BbsI-BbsI	64-174	111
4	BbsI-(RightEnd)	536-627	92
5	BbsI-BbsI	198-257	60
6	BbsI-BbsI	17-41	25
7	BbsI-BbsI	42-63	22
8	(LeftEnd)-BbsI	1-16	16
9	BbsI-BbsI	182-197	16
10	BbsI-BbsI	175-181	7

#	Ends	Coordinates	Length (bp)
1	(LeftEnd)-BbsI	1-16	16
2	BbsI-BbsI	17-41	25
3	BbsI-BbsI	42-63	22
4	BbsI-BbsI	64-174	111
5	BbsI-BbsI	175-181	7
6	BbsI-BbsI	182-197	16
7	BbsI-BbsI	198-257	60
8	BbsI-BbsI	258-399	142
9	BbsI-BbsI	400-535	136
10	BbsI-(RightEnd)	536-627	92

Figure 2.4: *DNA Fragments for the Bbs-I enzyme cutting example of Figure 2.3*
In the left table, fragments are sorted by length (distance in nucleotides to the next fragment)
In the right table, fragments are sorted by the coordinates (position) where the fragment is found in the DNA sequence.
Screenshot source : NEB.com

UC-2 Compare two DNA sequences (TO BE FURTHER DEFINED)

With 2 DNA sequences, given as an input to the system by the user, the system needs to compute a percentage of similarities between the two sequences.

For this operation, the system needs to consider the different DNA variations types (addition, removal, substitution, etc.) between the two DNA sequences. There is a reference to a DNA variant-oriented storage format in the master thesis of Sébastien Servoles.

(QUESTION: Should the DNA sequences should be used in the comparison, or just the position of the enzymes occurrences found by use case "UC-1 Locate an enzyme in a DNA sequence" ?)

(TODO: more details required from Dr. Trifiro)

3 Actors

This section reports all of the actors mentioned in the use-case model survey.

3.1 Lab Researcher

The Lab Researcher is the main end-user of the application developed in this project. It can be a lab technician, a doctor, etc.

4 Requirements

4.1 Functional Requirements

Functional requirements may be calculations, technical details, data manipulation and processing and other specific functionality that define what the Genetic Heterogeneity system is supposed to accomplish.

(TODO: in progress)

4.2 Non Functional Requirements

Nonfunctional requirements or quality attributes of a system are criteria that can be used to judge the operations of a system, rather than specific behaviours.

The following non functional requirements are classified using the Quint 2 model (see section 6.1 for more details about Quint 2).

(TODO: in progress)

5 Design Constraints

This section indicates any design decisions that have been mandated and that the Genetic Heterogeneity system must be adhered to.

(TODO: in progress)

6 Applicable Standards

This section references all the legal, quality, regulatory, and industry standards that apply to the Genetic Heterogeneity system.

6.1 Quint-2 Extended ISO 9126-1 Model of Sotware Quality

This model has been used to elicit and analyze the non-functional requirements (section 4.2) that applies to this project. The Quint 2 model classifies the quality attributes into 5 large classes : Functionality, Reliability, Efficiency, Maintainability, and Portability. At the time of writing, this model constituted the most complete reference available.

For more information, please visit :

http://web.archive.org/web/20100410024852/http://www.serc.nl/quint-book/index.htm

http://www.sqa.net/iso9126.html

Glossary

This section lists the definitions, acronyms, abbreviations or company-specific shorthands that are necessary for a contextual understanding of this document and the application.

(TODO: insert accurate definition)

- Enzyme
- Base pair direction (5' , 3')
- Nucleotide
- DNA sequence : in this text, we mean = whole genome, chromosome, or other DNA sub-sequence
- what is DNA cut ?
- Strand