

# Scipio Documentation

## Contents

- General remarks
- Evaluating the results in the YAML-file
- Yaml2log: evaluating the results in human readable format
- Types of discrepancies

## General remarks

The WebScipio interface offers the evaluation of the results at various positions. First, and most prominent, the overall result is given as complete or incomplete. Below the button Search Details all information regarding the search (e.g. target file, target name to which the query was mapped) are given as well as some overall statistics. The scheme of the gene structure outlines most of the discrepancies like gaps, frame-shifts, mismatches, and questionable introns (intron?). The Result tab "Evaluation" lists all the discrepancies (if there are any; see also below) between the query sequence and the sequence found in the target genome. It provides their exact position, it provides additional information about the type of discrepancy (via tooltips on the type symbol), and provides further information about the possible incompleteness of the search like the missing of a stop-codon. The Alignment tab provides a detailed view on the gene and shows an easily readable alignment of the query sequence with the translated target sequence and the target sequence itself.

The following descriptions are especially useful if you use the Scipio command-line tool while this information is provided by WebScipio directly.

## Evaluating the results in the YAML-file

In the YAML-file, the "status" of a hit can be one of the following:

auto	the complete query is correctly matched by the hit
partial	part of the query is correctly matched by the hit
incomplete	one of the cases "A" to "E" occurred
(manual	the hit has been edited manually)

In the Log-file, the "status" of a query can be:

complete	the query is matched completely by one single hit or several partial hits
incomplete	one of the cases "A" to "E" occurred

## The various cases causing status: !incomplete

A	!missing stopcodon	There is no stop codon in the genomic DNA after the last amino acid of the query.
B	!bad intron	At least one of the introns does not show appropriate 5' and/or 3' splice sites.
C	!mismatches	At least one of the amino acids of the query does not match the translation of the genomic DNA.
D	!sequence shift	Additional or missing bases have been identified that would lead to frame shifts during translation. Those are most probable due to sequencing/assembly problems, but might also hint to the existence of pseudogenes.
E.1	!gap to querystart/queryend	There are unmatched aminoacids at the N-/C- terminus of the query: the first hit for a query doesn't start with the first amino acid of the query (the last hit for a query doesn't end with the last amino acid).
E.2	!gap to previous/next hit	There are unmatched amino acids between hits for the same query on different targets.
E.3	!gap	There are unmatched amino acids between two exons of a single hit.

## Yaml2log: evaluating the results in human readable format

The yaml2log script converts YAML files into an easily readable log file with summary information about the results and clearly arranged sequence alignments (Figure). This file is parsed by WebScipio and the various sections can be accessed in the Result Tabs.

```
### yaml header:
### Scipio vl.4 (20100804-unreleased) output
# query file      /tmp/y46lIyL7tLE8.fasta
# target file     /genomes/genomes_ncbi/Fusarium_oxysporum_f__sp__lycopersici_4286_v1_c
                  hromosome.fasta
# BLAT output     /tmp/sHoehiWD6Xrz.psl
# Timestamp       Fri Aug 20 09:32:59 2010

#####
query           Protein
status          incomplete
-----
mismatches and sequence shifts:
  posn: exp.aa:  found in dna:    type:    example case:
    17:  Q      N      [aac]      M        1, mismatch
   259:  X      EX      [gaa t]    F:l,U,+  2,7, undetermined query /
                                      frameshift (+1) target/query
#####
ID              1
status          ! incomplete
```

```

reason          ! sequence shift
                ! mismatches
query length    551
introns at      95
target          gi|213958556|gb|CM000593.1| Fusarium oxysporum f. sp. lycopersici
                4286 chromosome 5, whole genome shotgun sequence
target length   4914260
strand          +
no of exons     2
target location 4050103..4051807
introns at      4050387..4050437
hit length      551
- matches       ! 549
- mismatches    ! 1
- undeterm.     ! 1
insertions      ! 1
identity        ! 99.6%
score           0.995
-----
upstream
  agtatgta ttgtacaccg tttatcagcc catgggactc tccgatccga tctatcggag catctcga      4048168
  acaccatc caccgaagct agtaggagca ttaccttagg ctttggtcac agcttgcgca atacaagt      4048234
  caagttac agcaaagaga agttggtgtg tgtctctgca tgatgcatcg acagtcacac aacaccac      4048300

```

## Types of discrepancies

The yam2log script identifies cases from a list of alignment discrepancies and mismatches between query and target sequence that can result from sequencing/assembly errors (Figure). The simplest case is that amino acids differ (cases 1 to 3), or that they are missing in either the target or the query (cases 4 and 5). Sequencing/assembly errors may lead to additional or missing bases. These frameshifts are represented by an X in the translation corresponding to one or two nucleotides. The query sequence might have either been obtained from cDNA sources thus leading to a mismatch between query and translated target (cases 6, 8, 10, and 12), or the sequencing errors might have already been interpreted represented by an X in the query (cases 7, 9, 11, and 13). The target sequence might also contain in-frame stop codons (cases 14 to 17). These can be the result of sequencing errors or real stop codons as they appear in pseudogenes. In all these cases, the stop codon is shown as an asterisk ('\*') in the translation.

1: mismatch

gDNA	AAA	ttt	GGG
translation	K	F	G
		X	
query	K	A	G

2: undetermined query

gDNA	AAA	ttt	GGG
translation	K	F	G
query	K	X	G

3: undetermined target

gDNA	AAA	nnn	GGG
translation	K	X	G
query	K	A	G

4: additional codon in target

gDNA	AAA	ttt	GGG
translation	K	F	G
query	K	-	G

5: unmatched query

gDNA	AAA	---	GGG
translation	K	-	G
query	K	A	G

6: frameshift (+1) target only

gDNA	AAA	t--	GGG
translation	K	X	G
query	K	-	G

7: frameshift (+1) target/query

gDNA	AAA	t--	GGG
translation	K	X	G
query	X	-	G

8: frameshift (+2) target only

gDNA	AAA	tt-	GGG
translation	K	X	G
query	K	-	G

9: frameshift (+2) target/query

gDNA	AAA	tt-	GGG
translation	K	X	G
query	X	-	G

10: frameshift (-2) target only

gDNA	AAA	t--	GGG
translation	K	X	G
query	K	A	G

11: frameshift (-2) target/query

gDNA	AAA	t--	GGG
translation	K	X	G
query	K	X	G

12: frameshift (-1) target only

gDNA	AAA	tt-	GGG
translation	K	X	G
query	K	A	G

13: frameshift (-1) target/query

gDNA	AAA	tt-	GGG
translation	K	X	G
query	K	X	G

14: stopcodon target/query

gDNA	AAA	tag	GGG
translation	K	*	G
query	K	*	G

15: stopcodon, target only

gDNA	AAA	tag	GGG
translation	K	*	G
		X	
query	K	D	G

16: stopcodon, undetermined query

gDNA	AAA	tag	GGG
translation	K	*	G
query	K	X	G

17: additional stopcodon

gDNA	AAA	tag	GGG
translation	K	*	G
query	K	-	G