Background

Siva collaborated in an interesting paper that looked for establish the phylogenetic relationships of NCR Legumes.

They used all the publicly Legumes genomes as well as the RNA-seq data from the plants below:

Clade	Species
IRLC	Medicago truncatula
	Medicago sativa
	Pisum sativium
	Cicer arietinum
	Trifolium pratense
	Melilotus officinalis
Dalbergioids	Arachis hypogaea
	Aeschynomene evenia
Robinoids	Lotus japonicus
Genistoids	Lupinus albus
	Lupinus luteus
	Lupinus mariae-josephae
Milletioids	Cajanus cajan
	Phaseolus vulgaris
	Vigna angularis
	Glycine max
Indigoferoids	Indigofera argentea

Objective

Predict NCR peptides on fenugreek and Sainfoin.
Because both genomes are not available in the NCBI, they were not predicted during Siva's collaboration. Fenugreek does not have a genome and Sainfoin does, however it is not published in the NCBI yet (as by 10-nov-2025).

Data

The Sainfoin genome was retrieved from its paper (as of 17-nov-2025 it is not in the NCBI)

The Fenugreek genome was assembled with Spades with the DNA data generated in this paper and saved under the SRA code ERR5639085

The Fenugreek transcriptome was assembled with RNA_Spades using the following public data: SRR14721915, SRR14721912, SRR14721913, SRR14721911, SRR14721914, SRR14721916

Methods

The plan was to apply SPADA a pipeline to predict NCRs.The pipeline is very old, therefore it was very difficult to install and run.

Here, for future learning I present the two approaches taken:

1. ncr_prediction_pre_spada
2. ncr_prediction_w_spada (This is the good method)

The workflows are for editing in Miro

TL;DR: The SPADA output was filtered by length (200aa) and number of Cysteine of mature peptides (>=4Cys). An the matured peptide classified using the HMM models for IRLC and Dalbergioids generated in this paper.

To validate the methods, Medicago truncatula was annotated and the curated list (ref_mtruncatula_NCRs.csv) of NCR (715 NCRs) from M. truncatula taken from Morphotype of bacteroids in different legumes correlates with the number and type of symbiotic NCR peptides used to confirm the results using MMseqs2 as aligner.

Question

For future: Is it possible to predict using only the signal?

Results

The following results are from SPADA:

All the metrics are for the filtered peptides.

Species	Number of NCRs	Number of NCRs (filtered)	sum_len	min_len	avg_len	max_len	Homology
M. Truncatula	1,205	940	89,214	42	94.9	227	CRPs_w_c
Sainfoin	1,804	1,173	135,010	39	115.1	225	CRPs_w_c
Fenugreek (transcriptome)	626	NAN	58,234	23	93	622	CRPs_w_c
Fenugreek (genome)	609	NAN	51,579	21	84.7	353	CRPs_w_c
Fenugreek all (no dups)	1,083	823	74,090	47	90	243	CRPs_w_c

The following results were for the alternative workflow (just for learn):

RBH=Reciprocal Best Hit

Species	Number of NCRs	RBH to M. truncatula Curated List
Medicago Truncatula	992	466
Sainfoin	1614	56
Fenugreek (transcriptome)	427	169

As you noticed the number of retrieved peptides was lower than the expected. Particularly, in M. truncatula NCR247 was not found. Fortunately, Jonathon was abled to fix the SPADA pipeline and I could learnt why I was not finding NCR 247 in my predictions: SPADA reported that a mix of different gene predictors cause differences in sensitivity. The best sensitivity was provided by the default: Augustus Evidence; GeneWise & SplicePredictor. **It turned out that the sensitivity was compromised because I was using only Augustus_Evidence and NCR247 was found by GeneWise;SplicePredictor!**

Curious fact

I tried to confirmed the NCRs using the proteome. Only 13 NCRs were detected (cov. and id. 0.9). Confirming that even 12 year after SPADA pipeline, current annotators are still not ready to annotate CRPs.