Bioinformatic analyses of CRISPR elements

Rolf Backofen

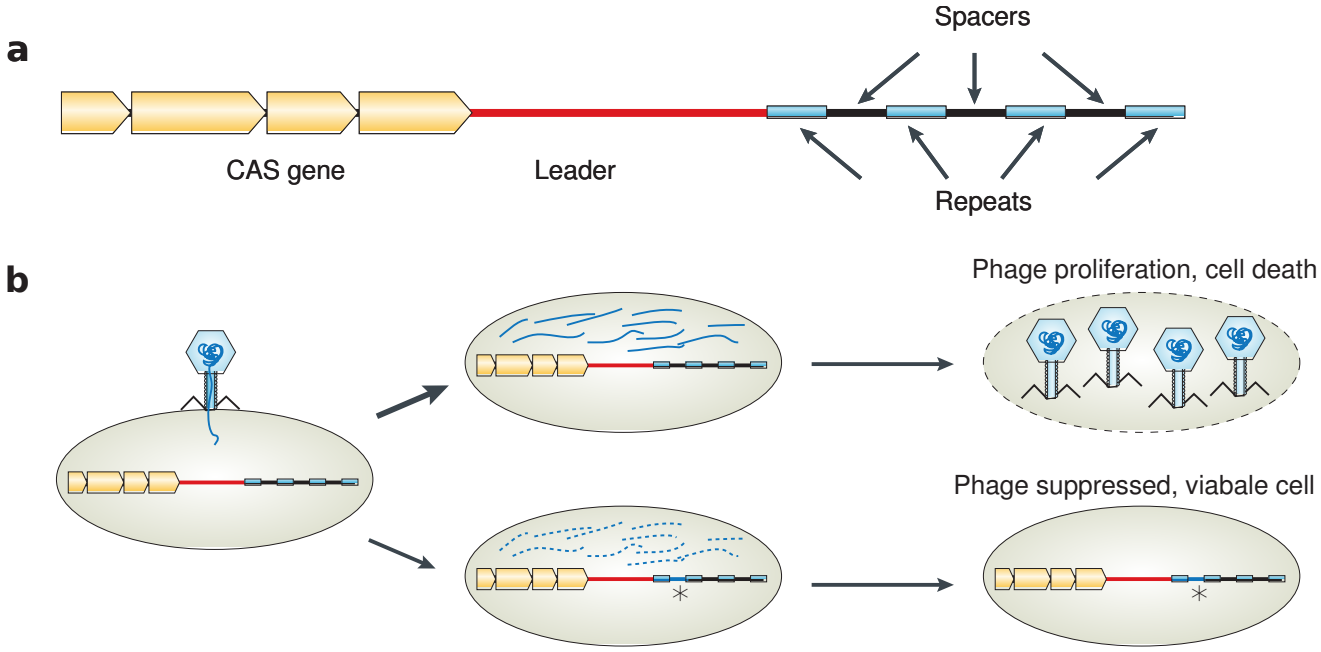
Lehrstuhl für Bioinformatik
Institut für Informatik

16. Februar 2012

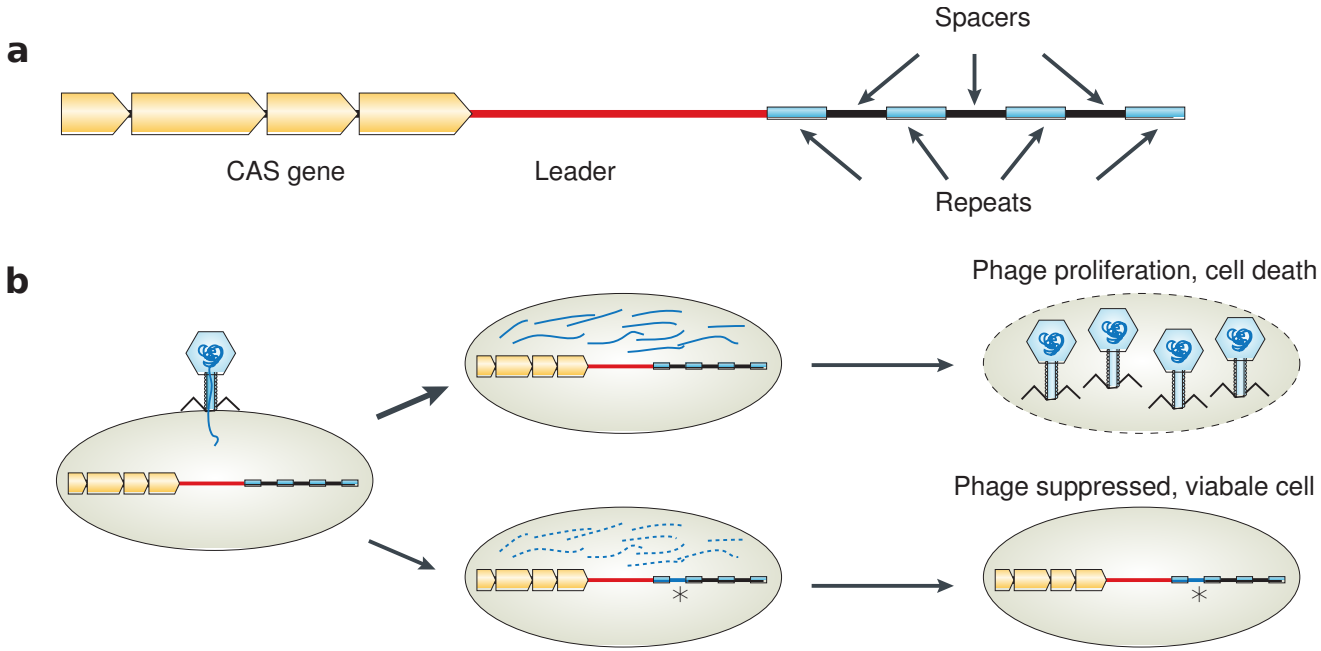
Overview:

- CRISPR Repeat Structure
- Analysis of Leader Sequences

CRISPR: Procaryotic Immune System

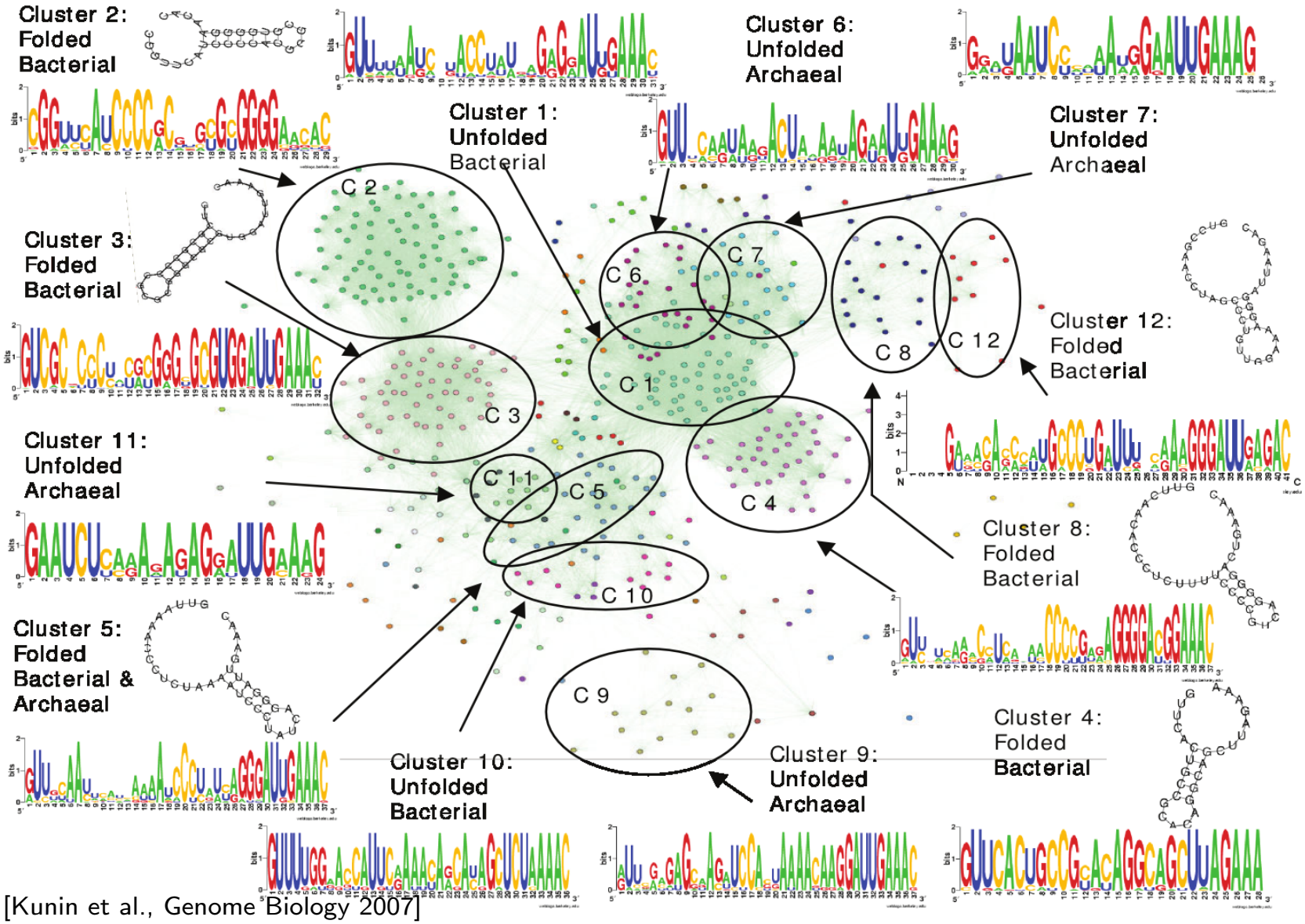


CRISPR: Procaryotic Immune System



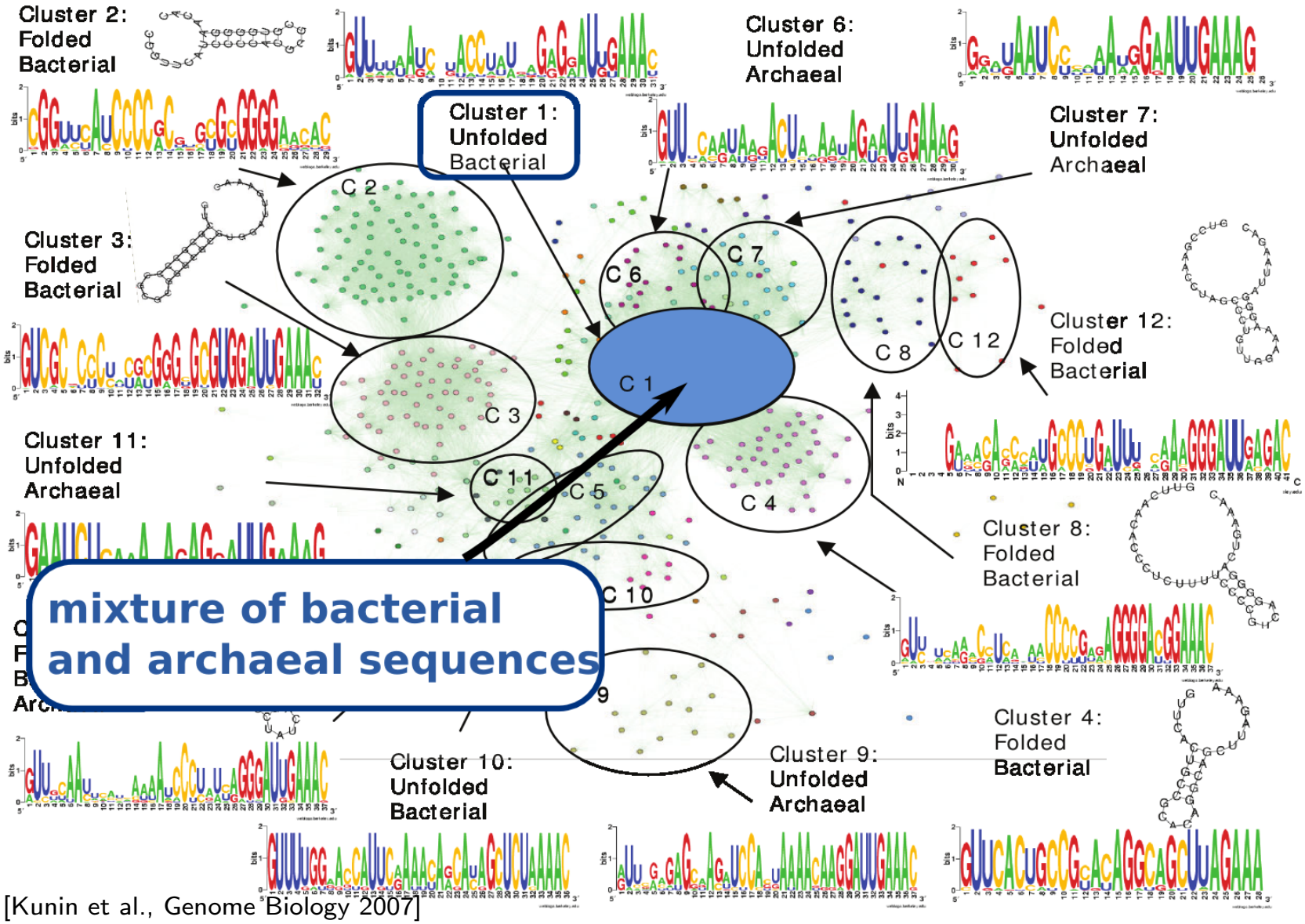
[Sorek et al. 2008]

Cluster Analysis of CRISPR Repeats



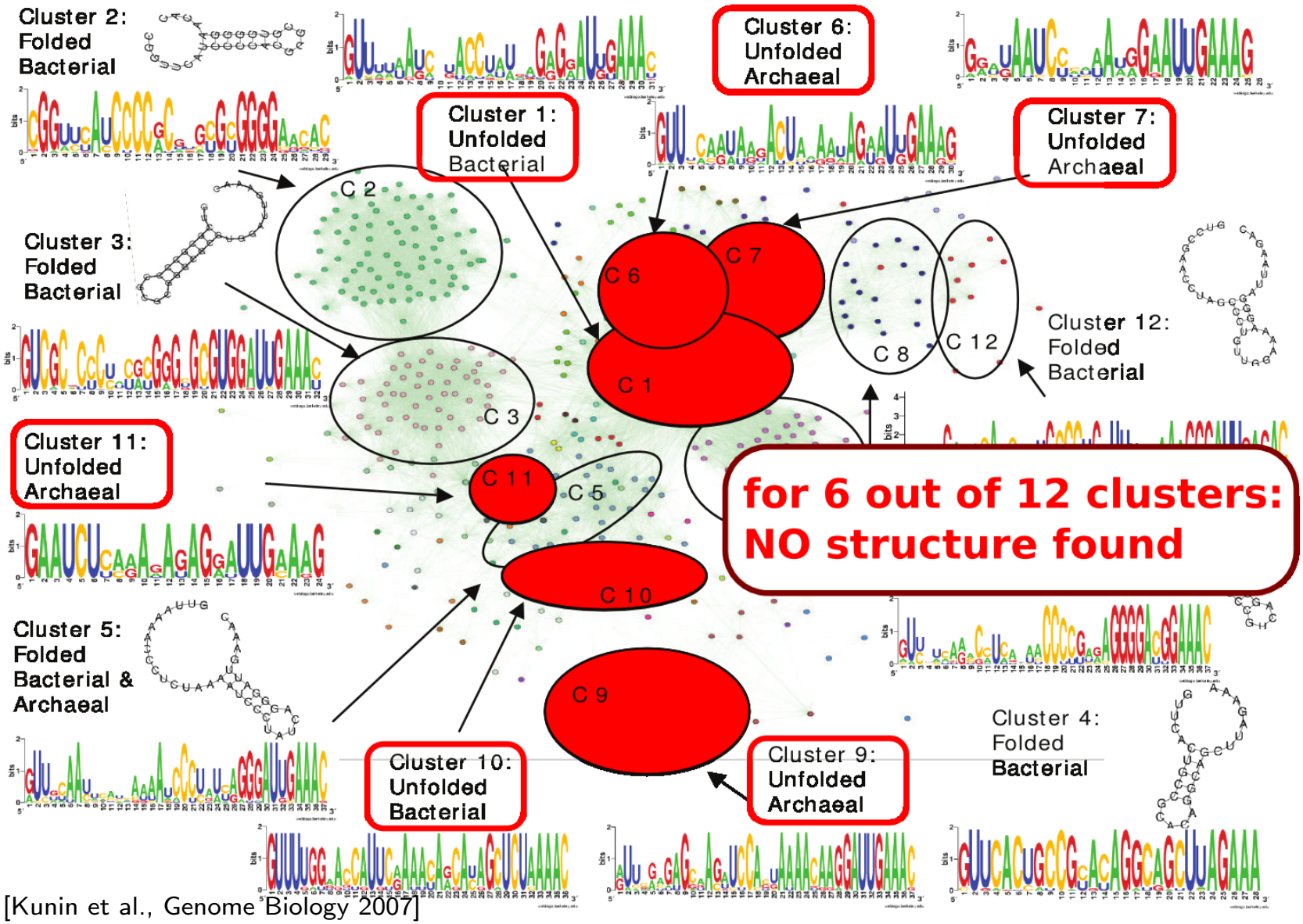
[Kunin et al., Genome Biology 2007]

Cluster Analysis of CRISPR Repeats



[Kunin et al., Genome Biology 2007]

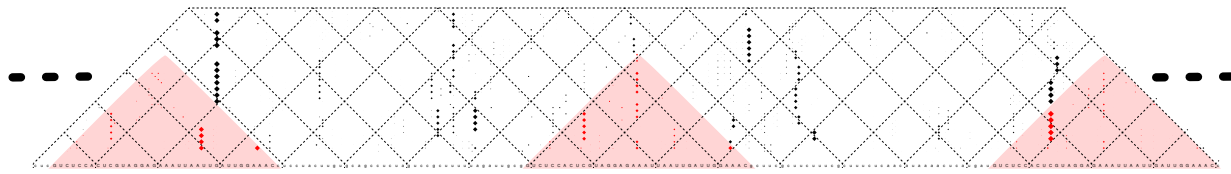
Cluster Analysis of CRISPR Repeats



[Kunin et al., Genome Biology 2007]

What is the Structure of a Repeat?

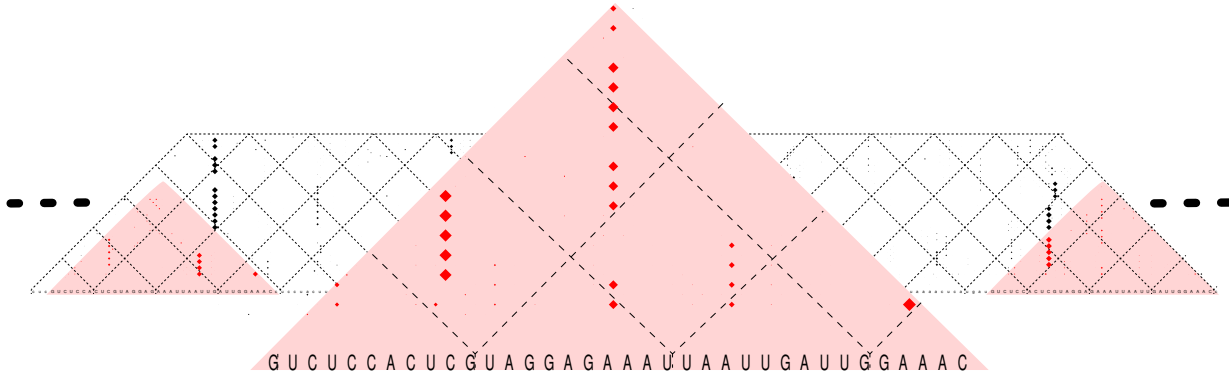
- example: 3 repeats of CRISPR array
- *problem: sub-optimal structures*



[Will et al: PLoS Comp. Bio 07]

What is the Structure of a Repeat?

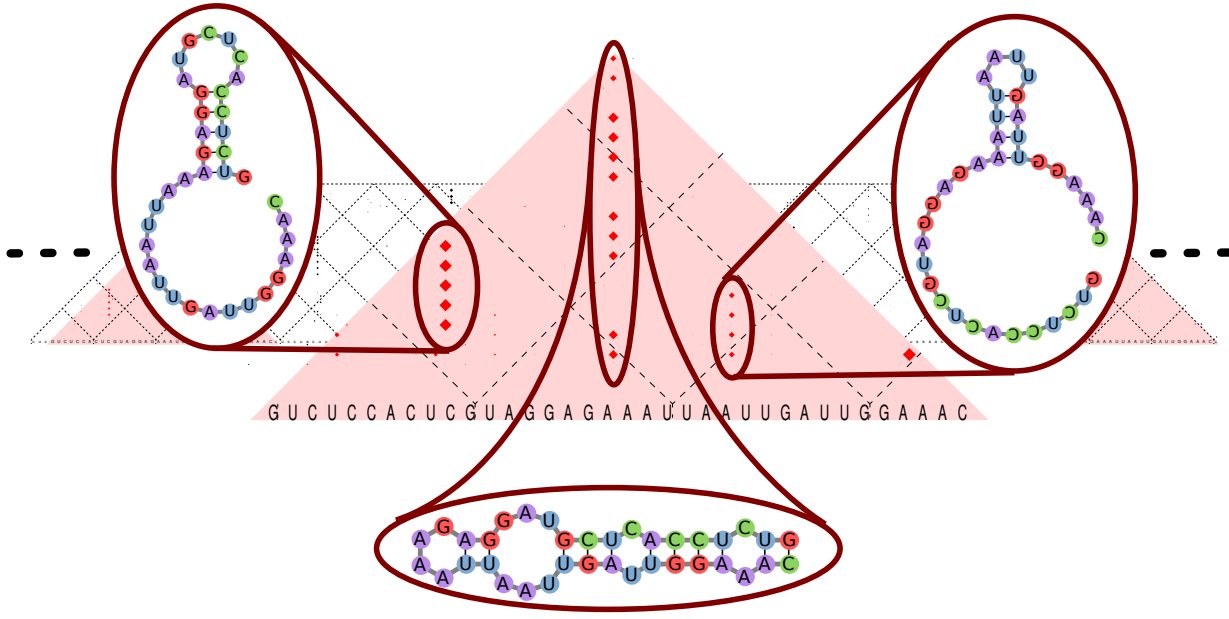
- example: 3 repeats of CRISPR array
- *problem: sub-optimal structures*



[Will et al: PLoS Comp. Bio 07]

What is the Structure of a Repeat?

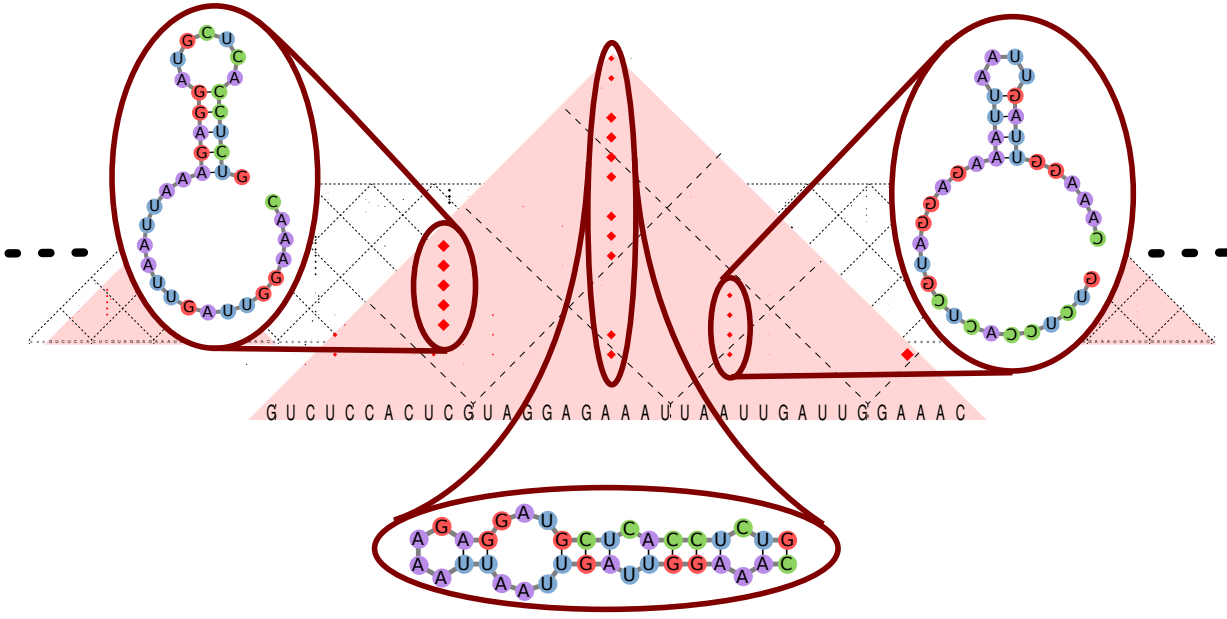
- example: 3 repeats of CRISPR array
- *problem: sub-optimal structures*



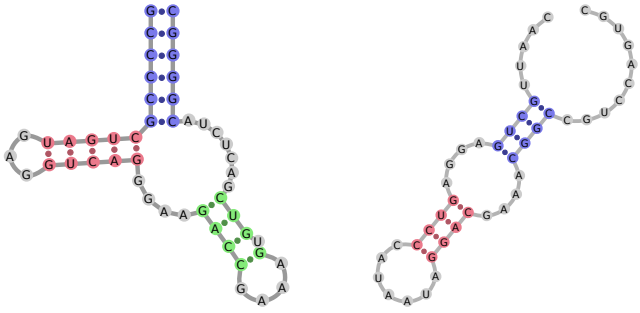
[Will et al: PLoS Comp. Bio 07]

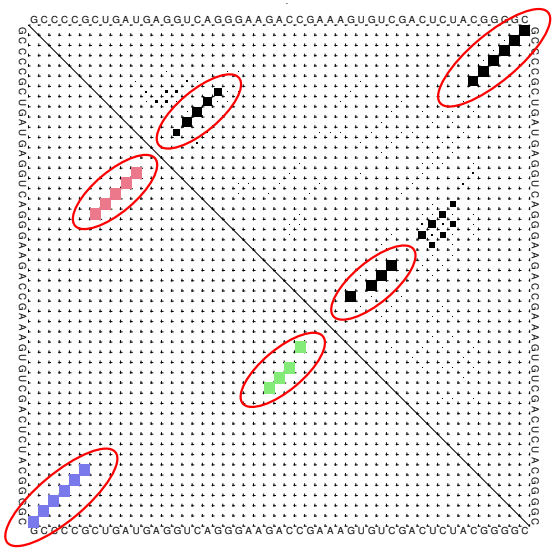
What is the Structure of a Repeat?

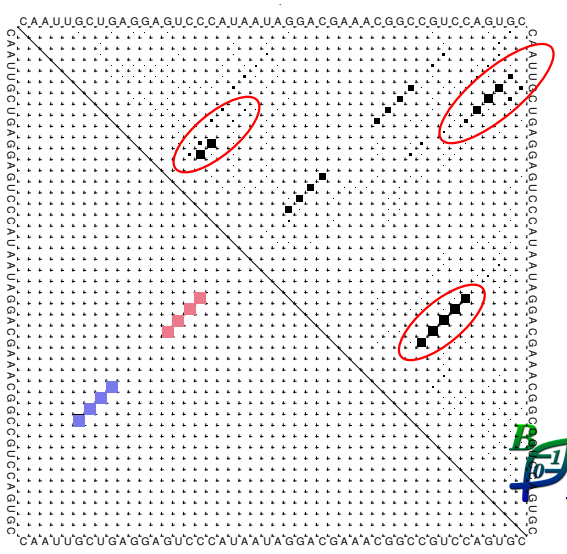
- example: 3 repeats of CRISPR array
- *problem: sub-optimal structures*



- LocaRNA: alignment of dotplots

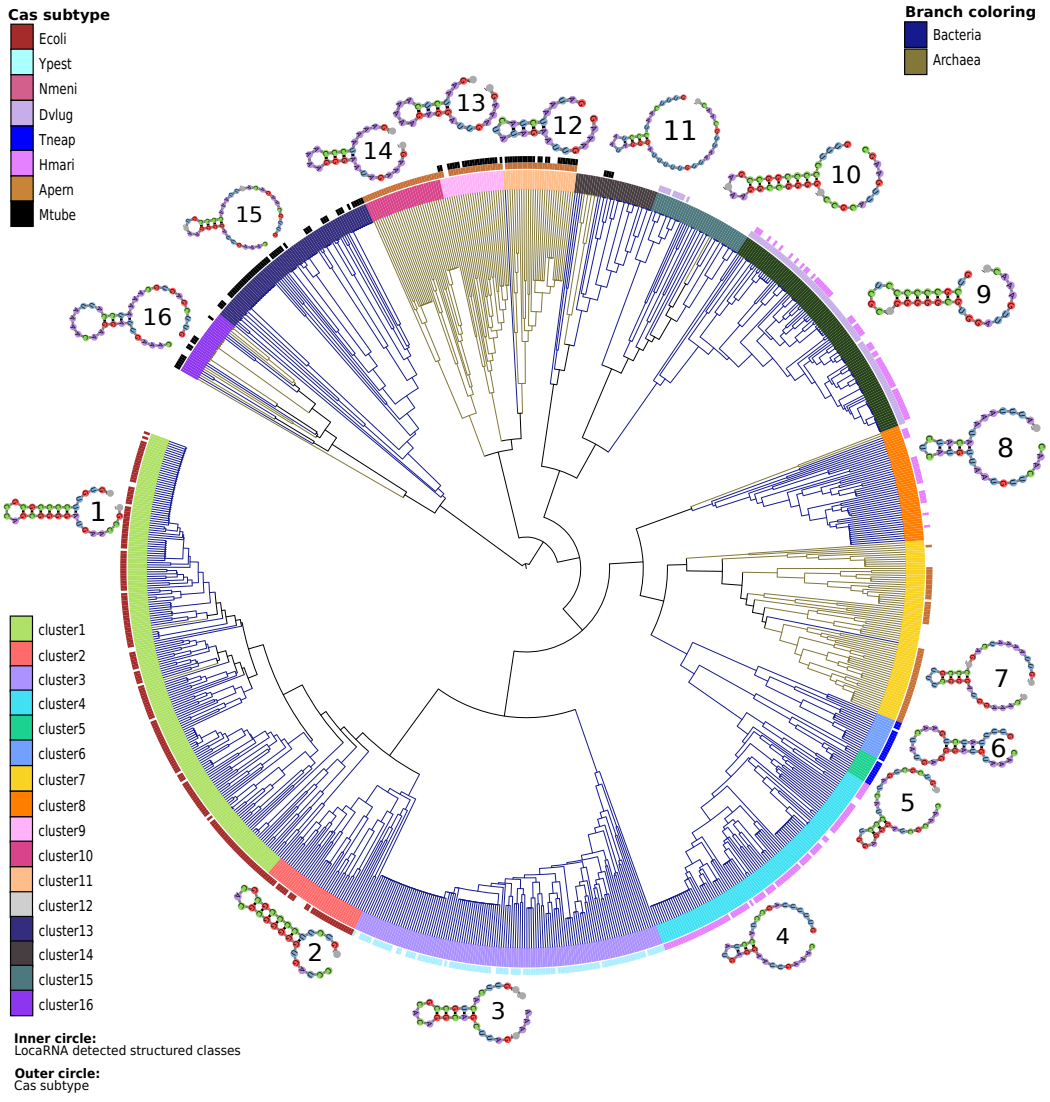






[Will et al: PLoS Comp. Bio 07]

Seq.-Struct. Clustering of CRISPR loci



CRIPSR:

- archaeal and bacterial genomes from CRISPI and CRISPdb
- CRT tool and CRISPRfinder
- total: 2020 crispr arrays in 670 genomes
- clustering with LocaRNA

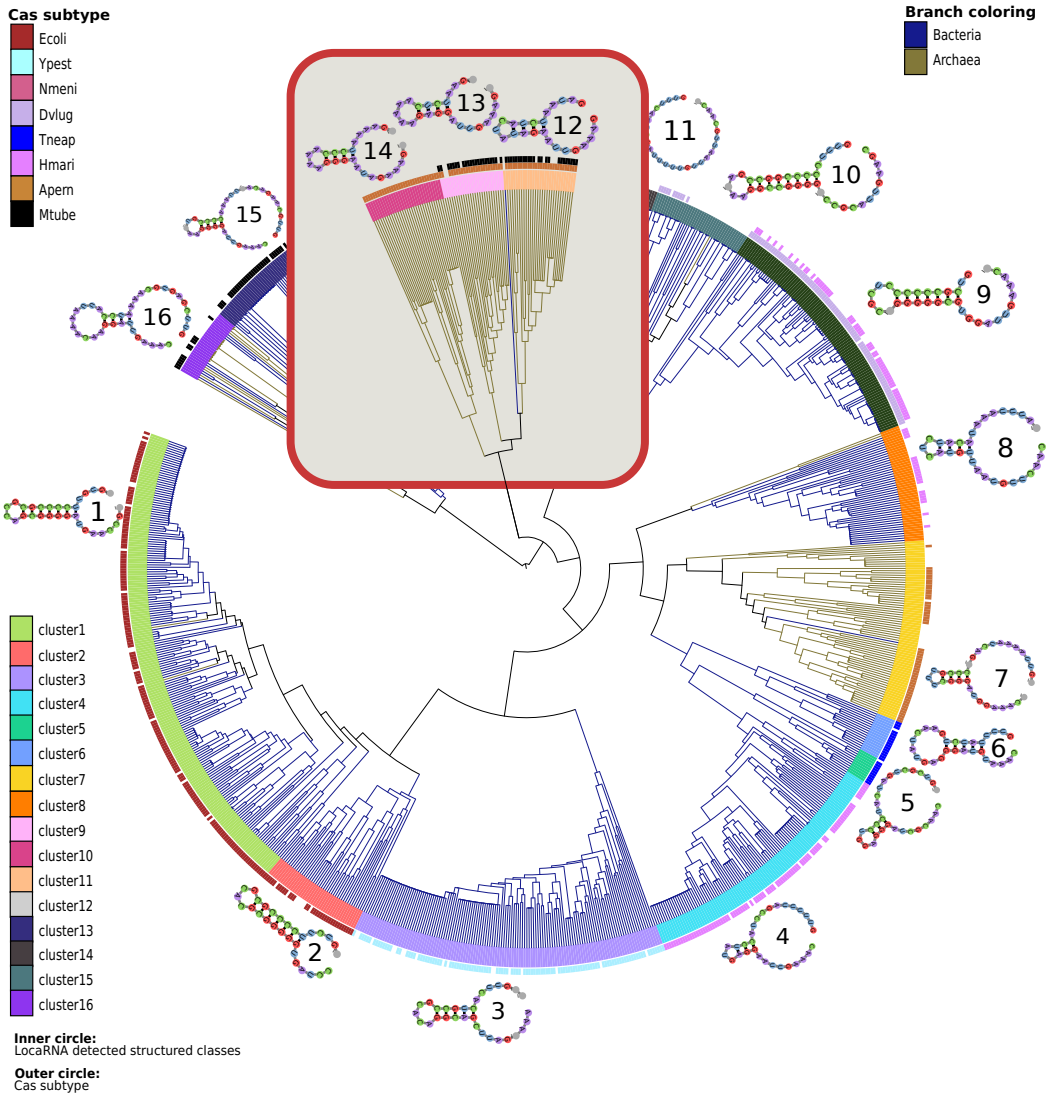
Cas-Genes:

- 45 CAS gene families (Haft et al.)
- search 20kb flanking regions

Results:

- Archaea and Bacteria in separated subtrees.
- perfect match with cas subtyping

Seq.-Struct. Clustering of CRISPR loci



CRIPSR:

- archaeal and bacterial genomes from CRISPI and CRISPdb
- CRT tool and CRISPRfinder
- total: 2020 crispr arrays in 670 genomes
- clustering with LocaRNA

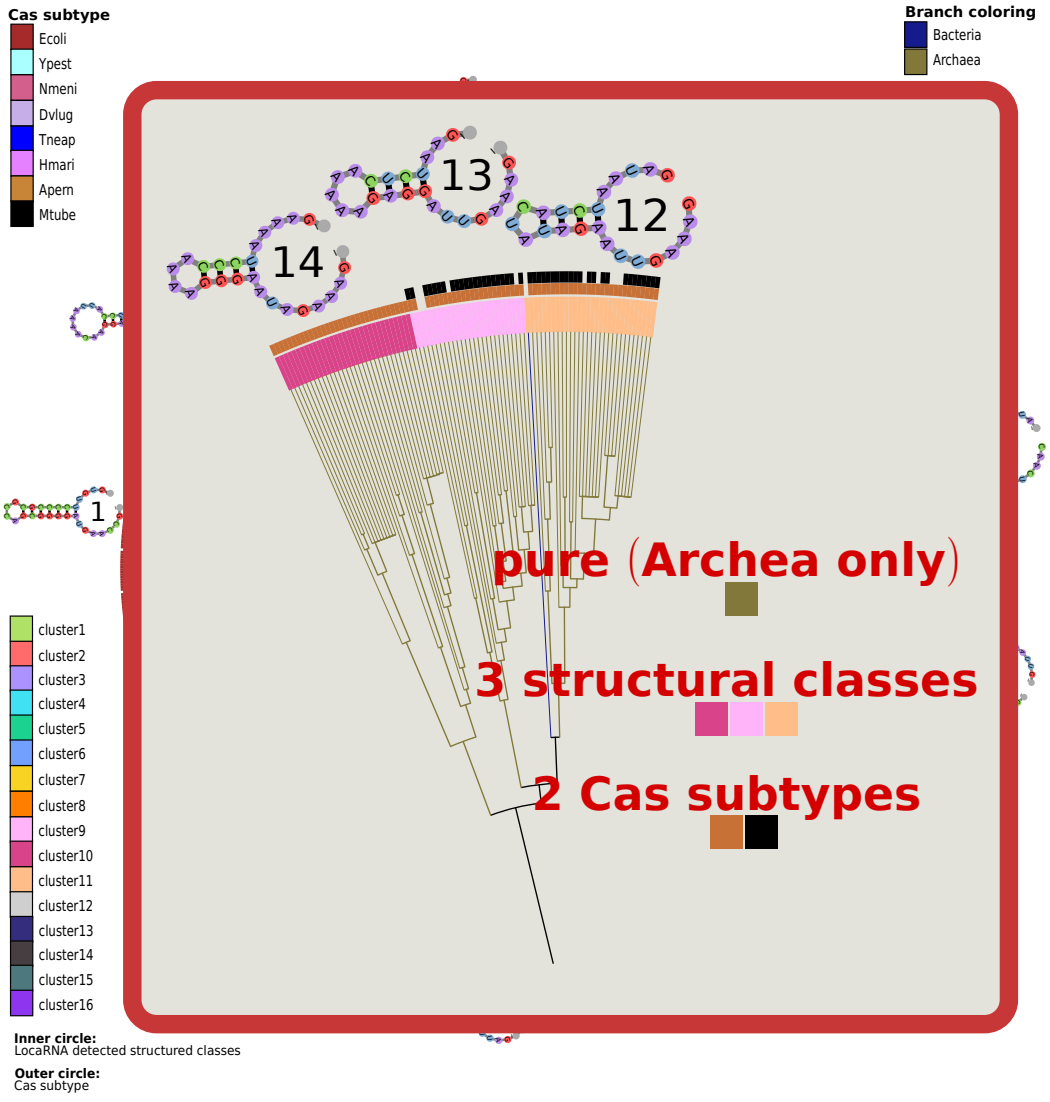
Cas-Genes:

- 45 CAS gene families (Haft et al.)
- search 20kb flanking regions

Results:

- Archaea and Bacteria in separated subtrees.
- perfect match with cas subtyping

Seq.-Struct. Clustering of CRISPR loci



CRIPSR:

- archaeal and bacterial genomes from CRISPI and CRISPdb
- CRT tool and CRISPRfinder
- total: 2020 crispr arrays in 670 genomes
- clustering with LocaRNA

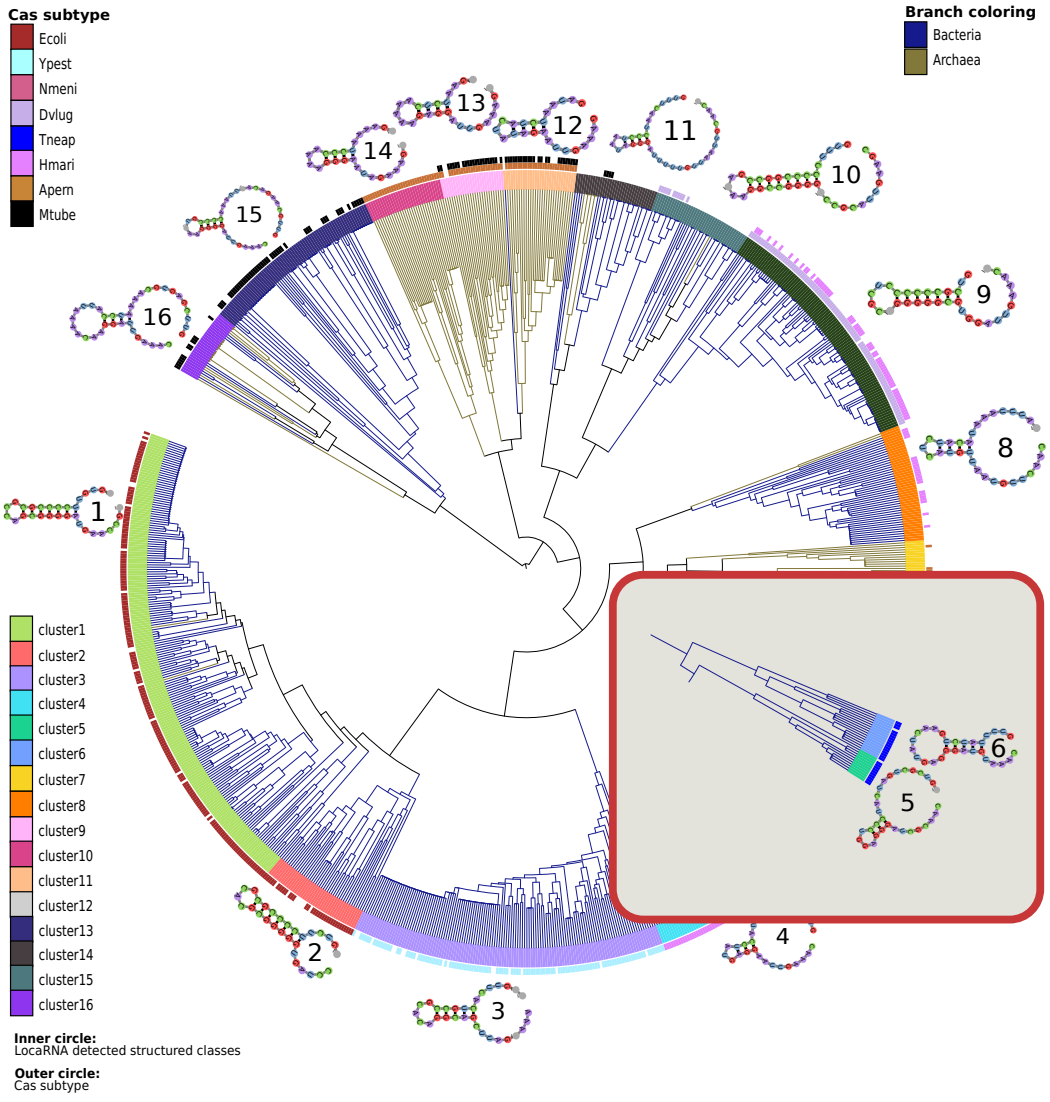
Cas-Genes:

- 45 CAS gene families (Haft et al.)
- search 20kb flanking regions

Results:

- Archaea and Bacteria in separated subtrees.
- perfect match with cas subtyping

Seq.-Struct. Clustering of CRISPR loci



CRIPSR:

- archaeal and bacterial genomes from CRISPI and CRISPdb
- CRT tool and CRISPRfinder
- total: 2020 crispr arrays in 670 genomes
- clustering with LocaRNA

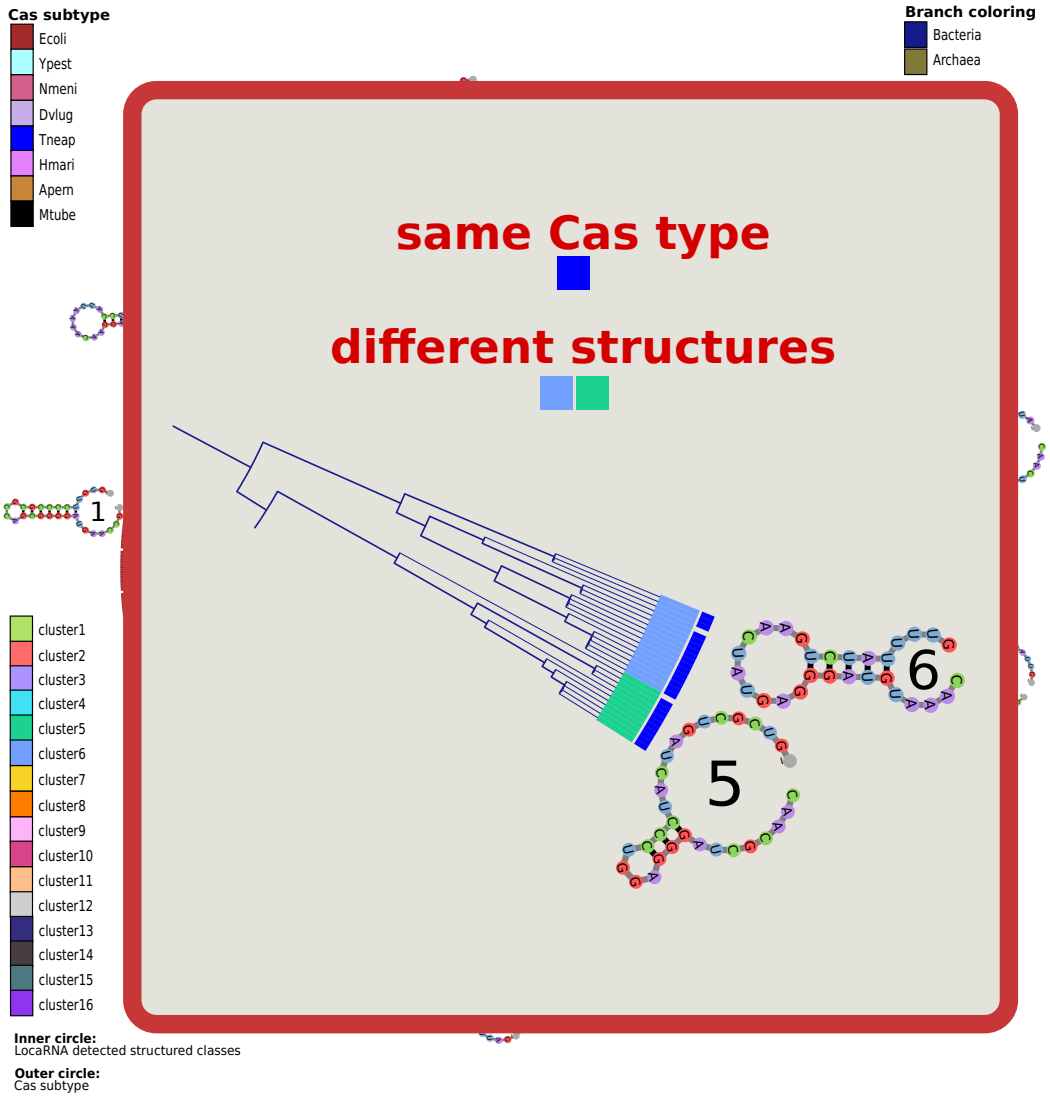
Cas-Genes:

- 45 CAS gene families (Haft et al.)
- search 20kb flanking regions

Results:

- Archaea and Bacteria in separated subtrees.
- perfect match with cas subtyping

Seq.-Struct. Clustering of CRISPR loci



CRIPSR:

- archaeal and bacterial genomes from CRISPI and CRISPdb
- CRT tool and CRISPRfinder
- total: 2020 crispr arrays in 670 genomes
- clustering with LocaRNA

Cas-Genes:

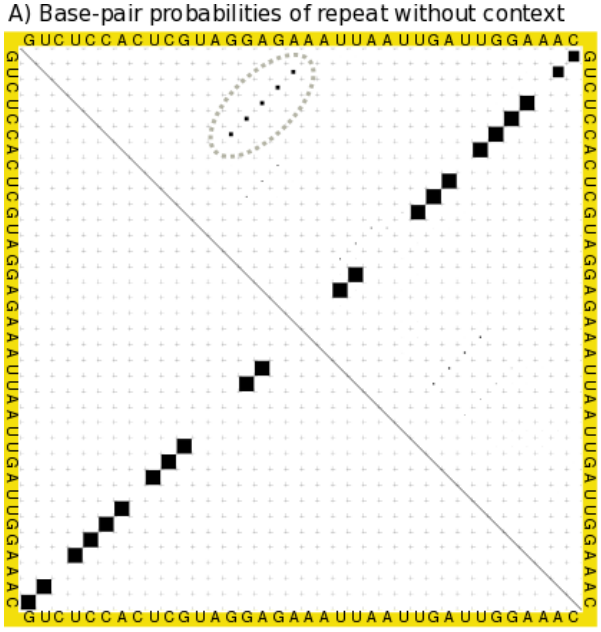
- 45 CAS gene families (Haft et al.)
- search 20kb flanking regions

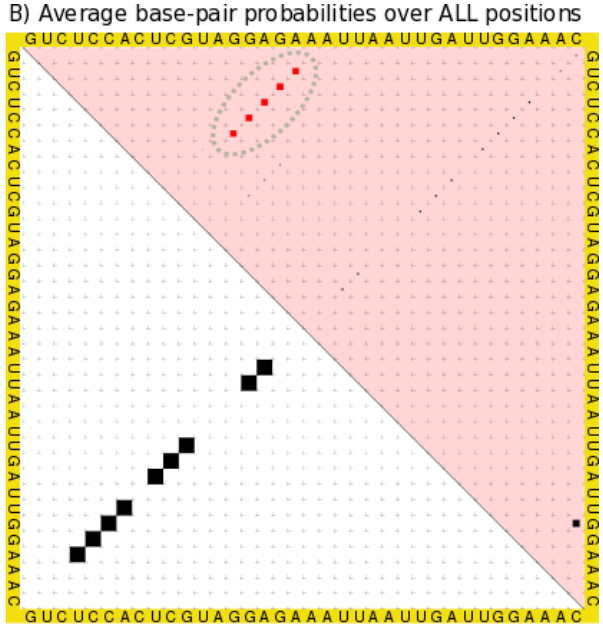
Results:

- Archaea and Bacteria in separated subtrees.
- perfect match with cas subtyping

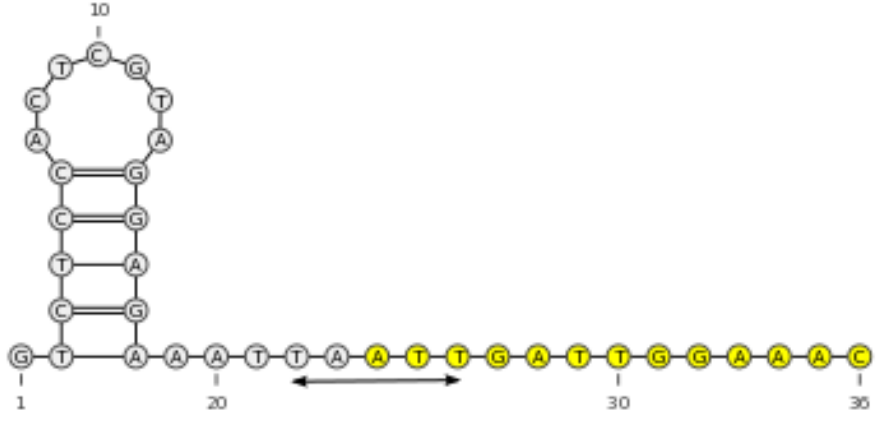
CRISPR3 from Synechocystis

- MFE structure as well as 5'/3' alternative structures
- 5' structure more dominant when folded in context of spacers



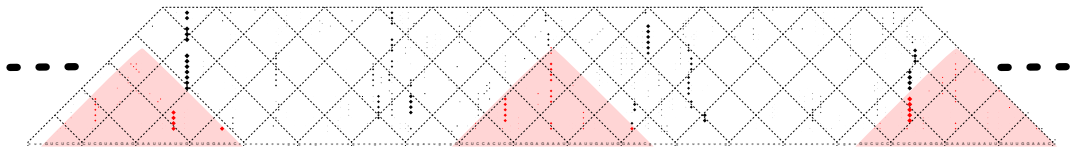


- resulting structure: agreement with conserved structure in cluster tree



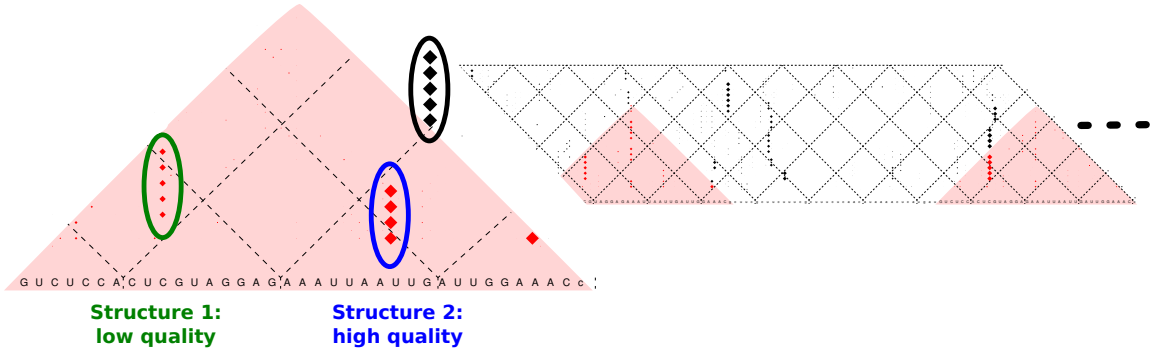
Which Structure is Correct?

- observation: different qualities $\Rightarrow$different processing order?



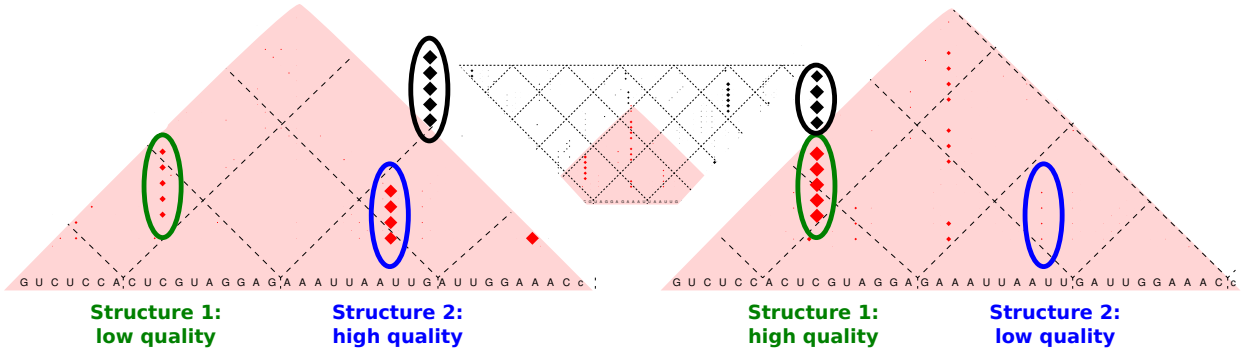
Which Structure is Correct?

- observation: different qualities ⇒different processing order?



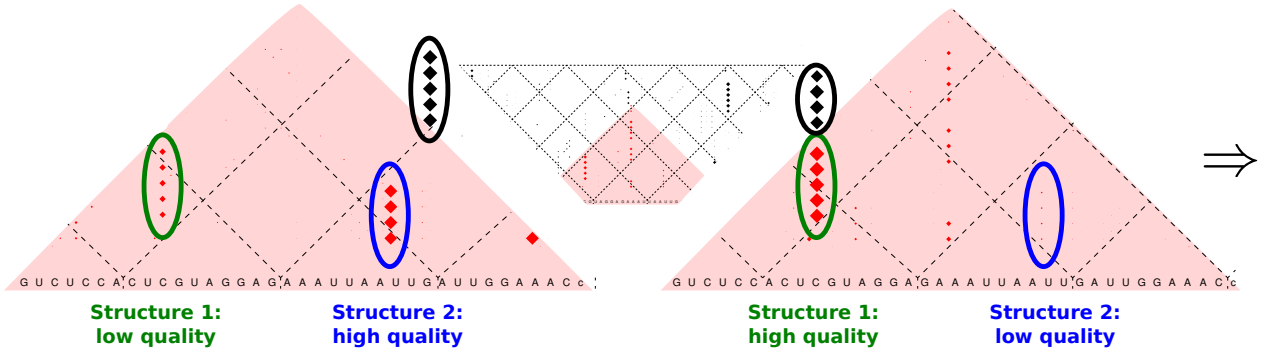
Which Structure is Correct?

- observation: different qualities ⇒different processing order?



Which Structure is Correct?

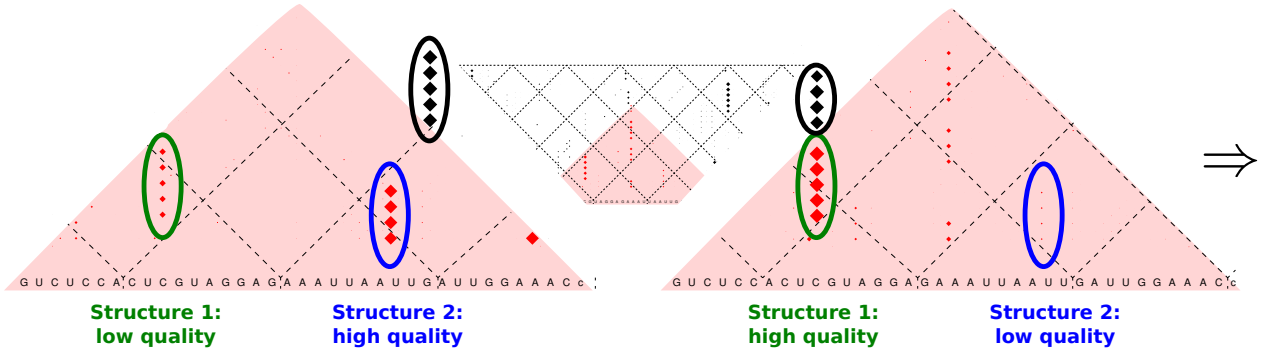
- observation: different qualities ⇒different processing order?



expected accuracy

Which Structure is Correct?

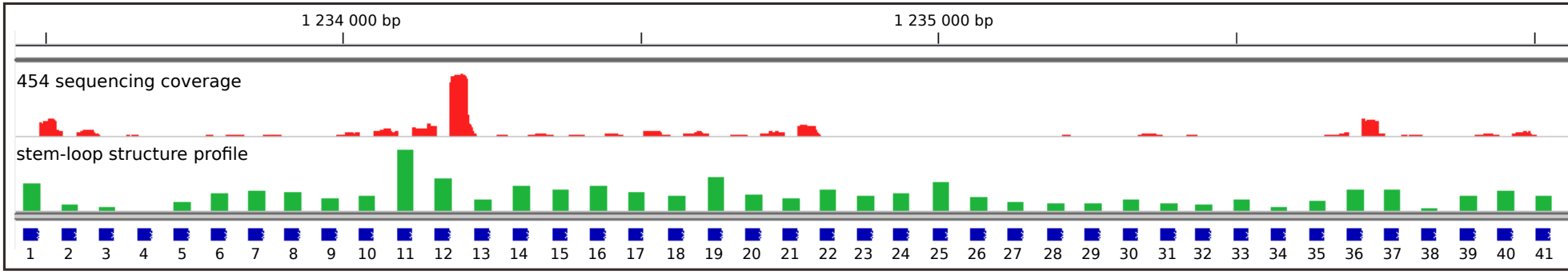
- observation: different qualities ⇒different processing order?



expected accuracy

- comparison with deep sequencing data

(*Sulfolobus solfataricus*)

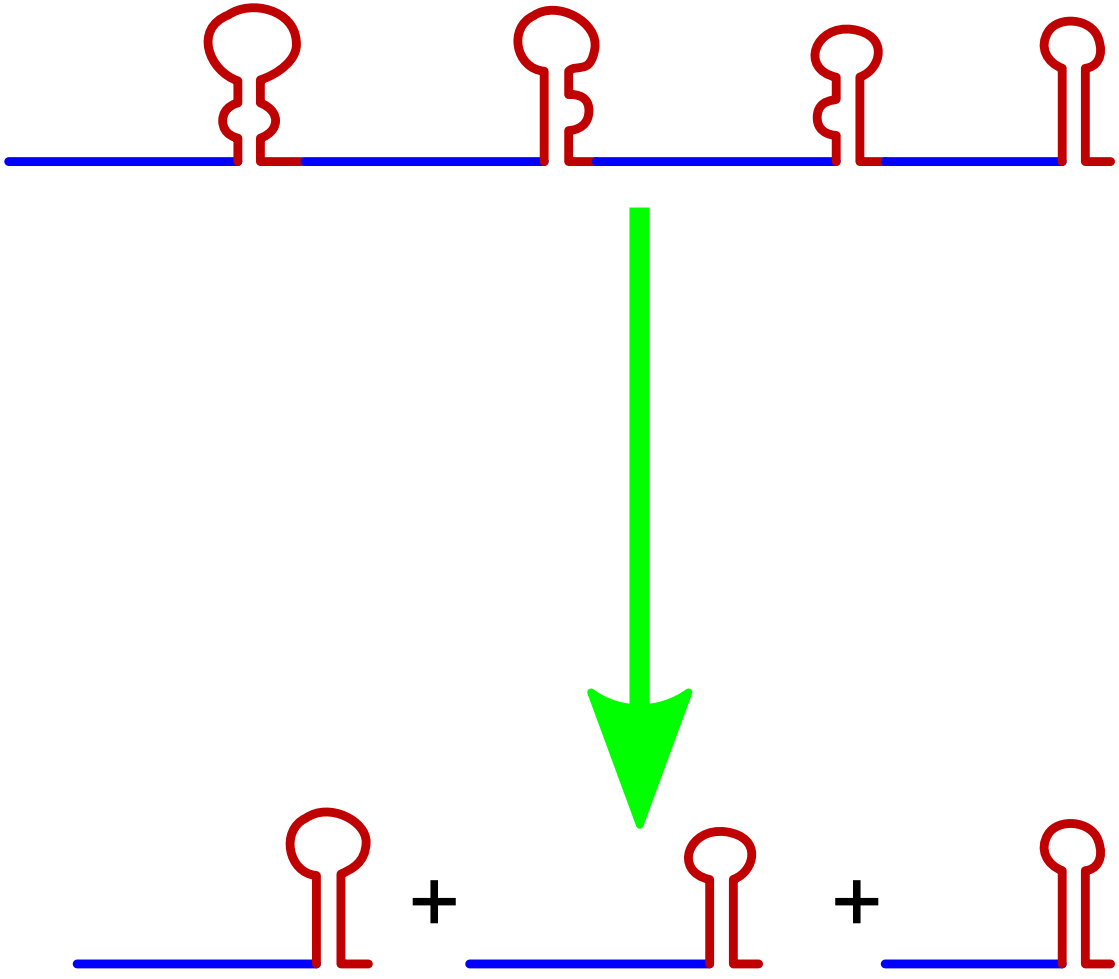


Problems so Far

- nice to find peaks
- required for statistical significance:
  correlation structure quality $\leftrightarrow$ *sequence reads*
- **However:** processing order makes troubles

Quantification and Processing Order

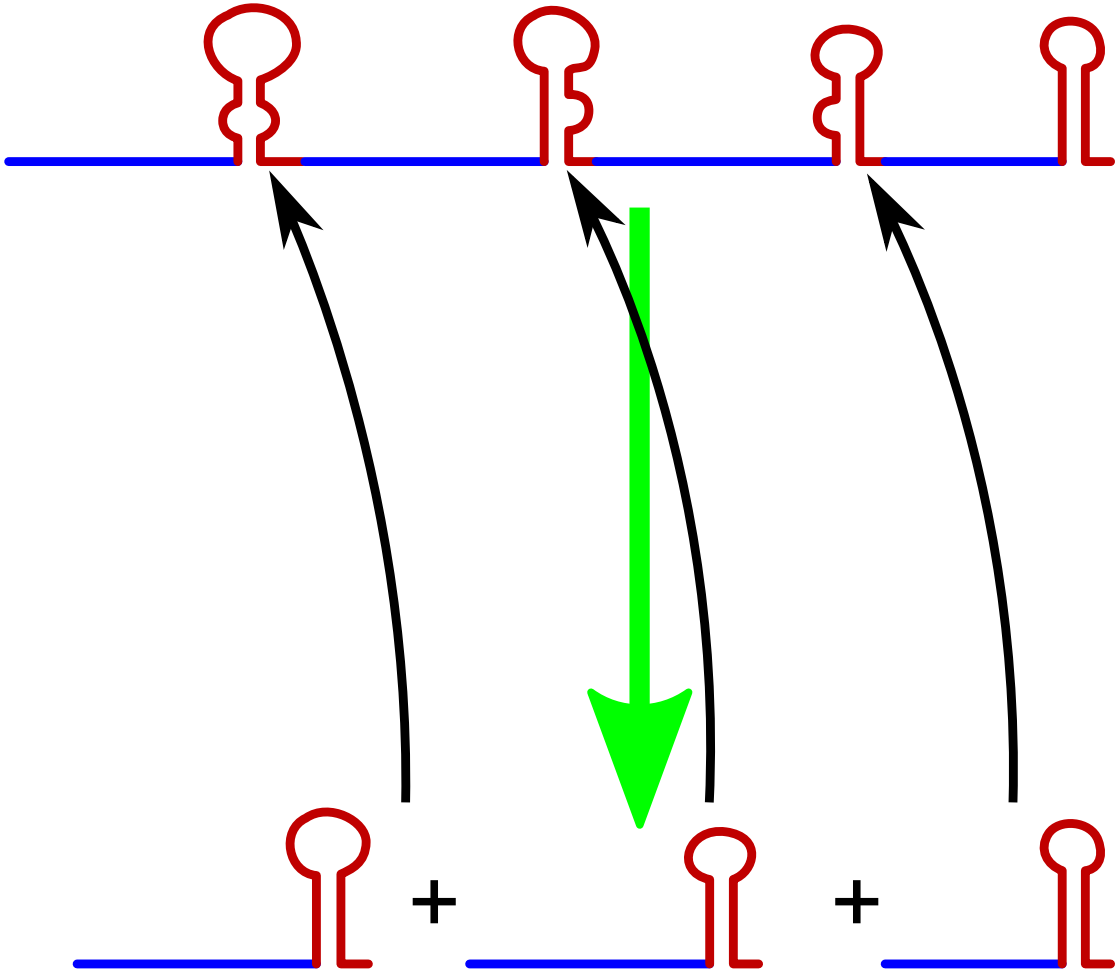
- question: how to correlate structure quality with sequencing reads?



Quantification and Processing Order

- question: how to correlate structure quality with sequencing reads?

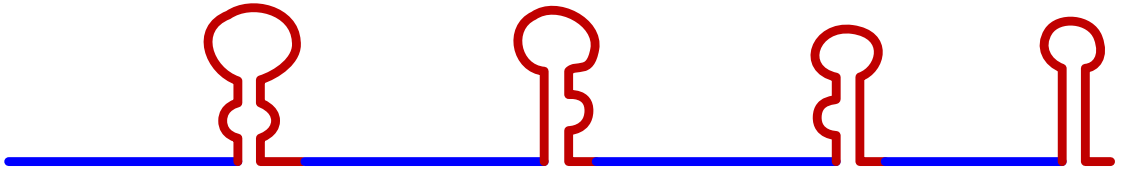
number of reads relates to structure quality of



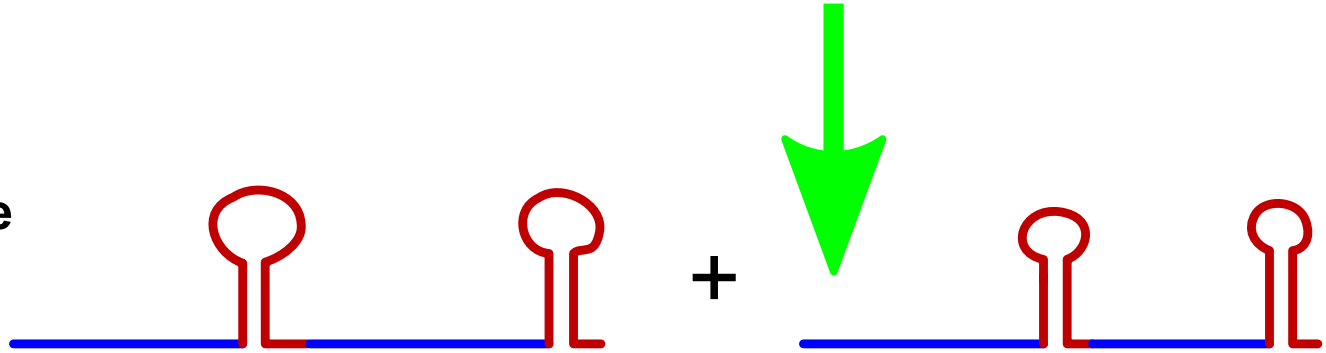
- could be measured with correlation

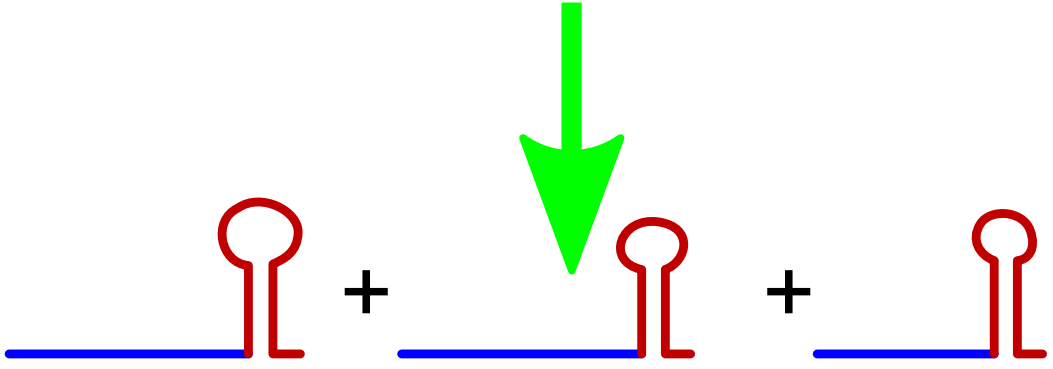
Quantification and Processing Order

- question: how to correlate structure quality with sequencing reads?



number of reads relates to structure quality of



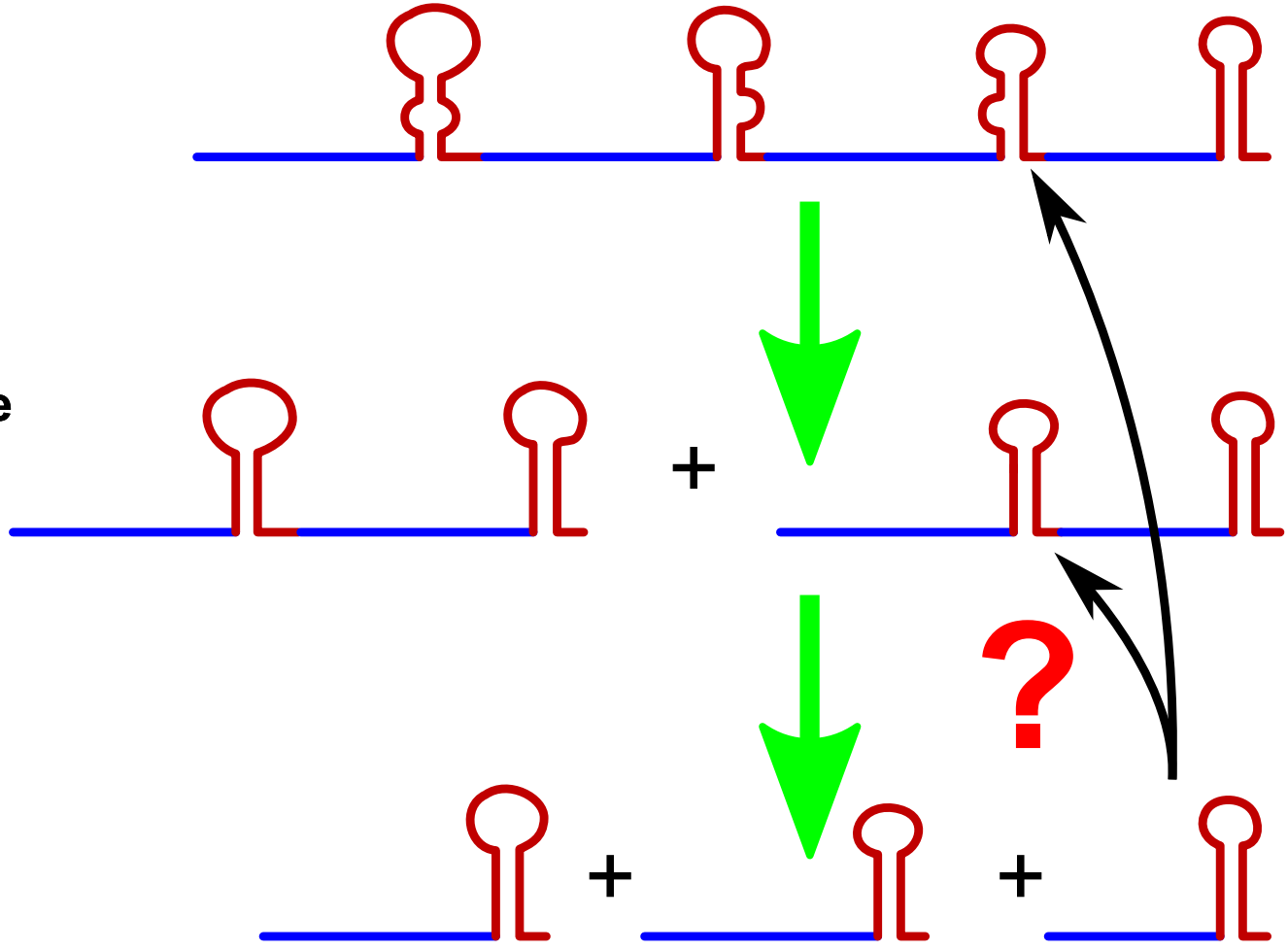


- could be measured with correlation

Quantification and Processing Order

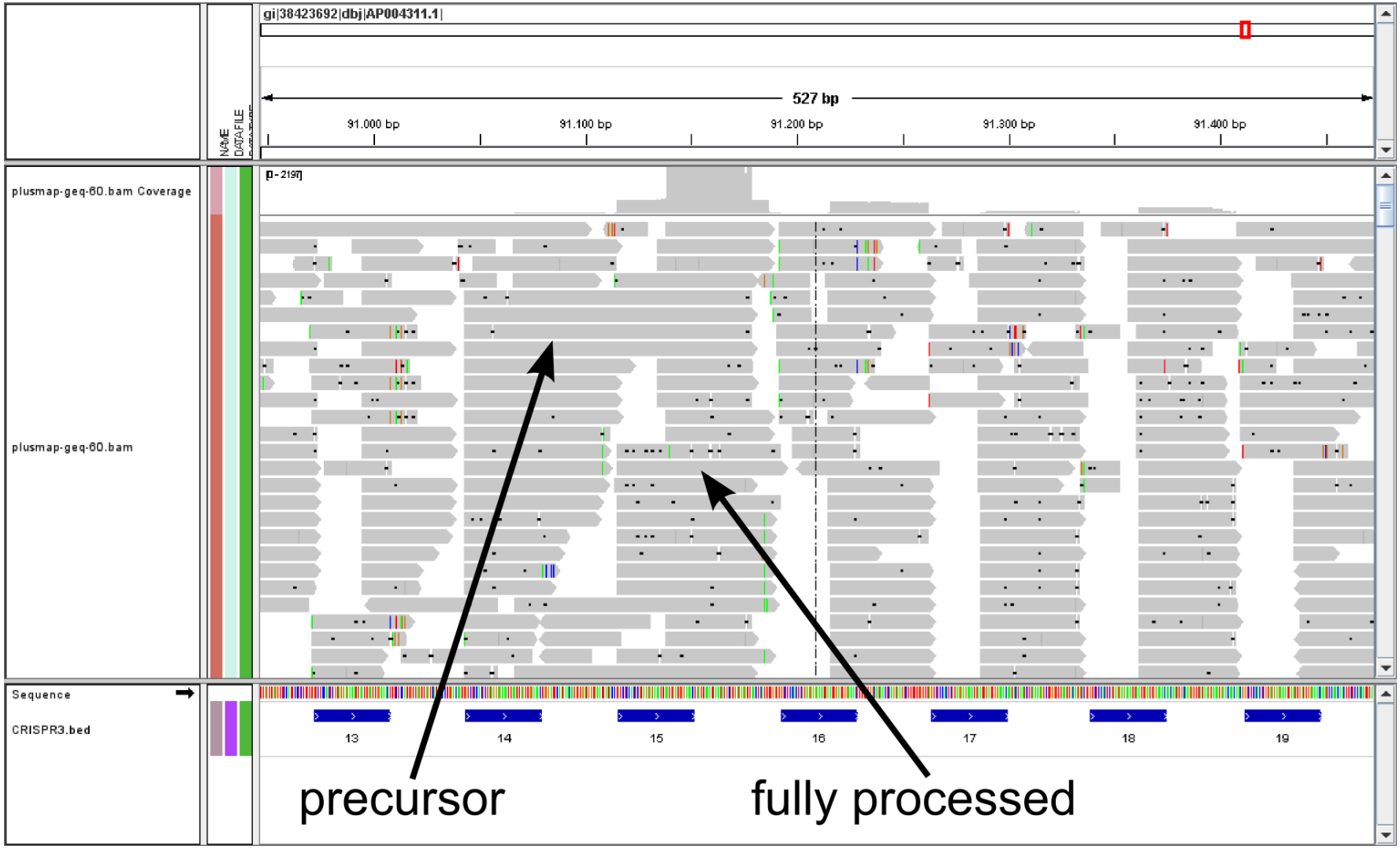
- question: how to correlate structure quality with sequencing reads?

number of reads relates to structure quality of which structure?



- could be measured with correlation
- now: correlation between what?

Example of Read Pattern



What Can be Done?

possible solution 1: different protocoll

- new sequencing data from Wolfgang Hess
- only longer sequences, longer than one repeat
- probably better reflection of first cleavage site

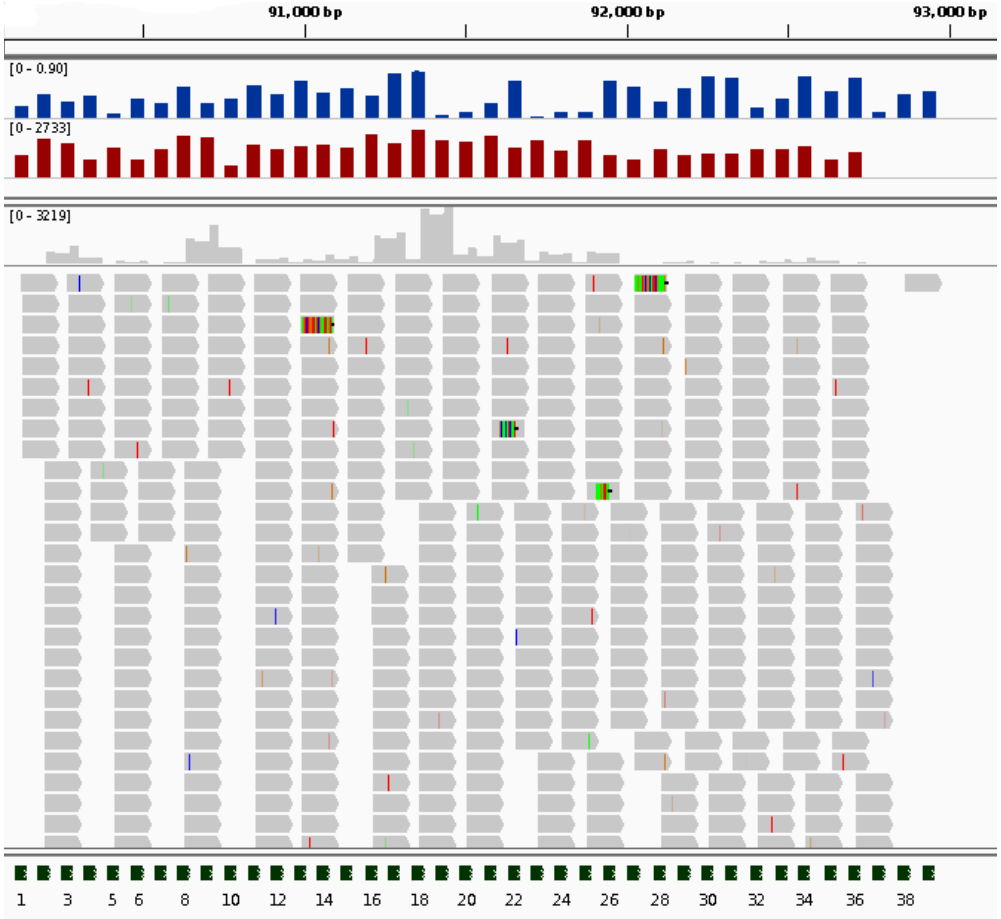
possible solution 2: different question

- what happens to fully processed spacers
  $\Rightarrow$ *investigation of degradation*

possible solution 3: design experiments (Lennart Randau)

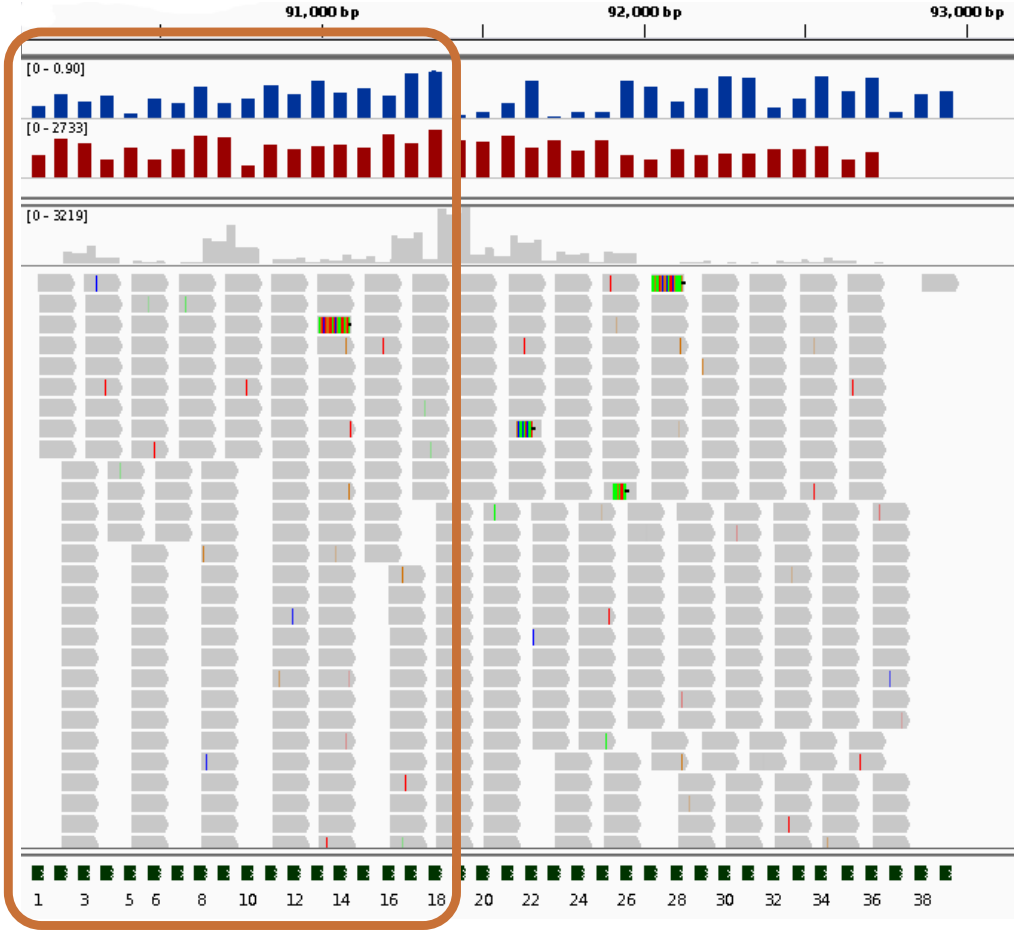
Solution 1: Different Protocol

CRISPR3 of Synechocystis sp. PCC 6803



Solution 1: Different Protocol

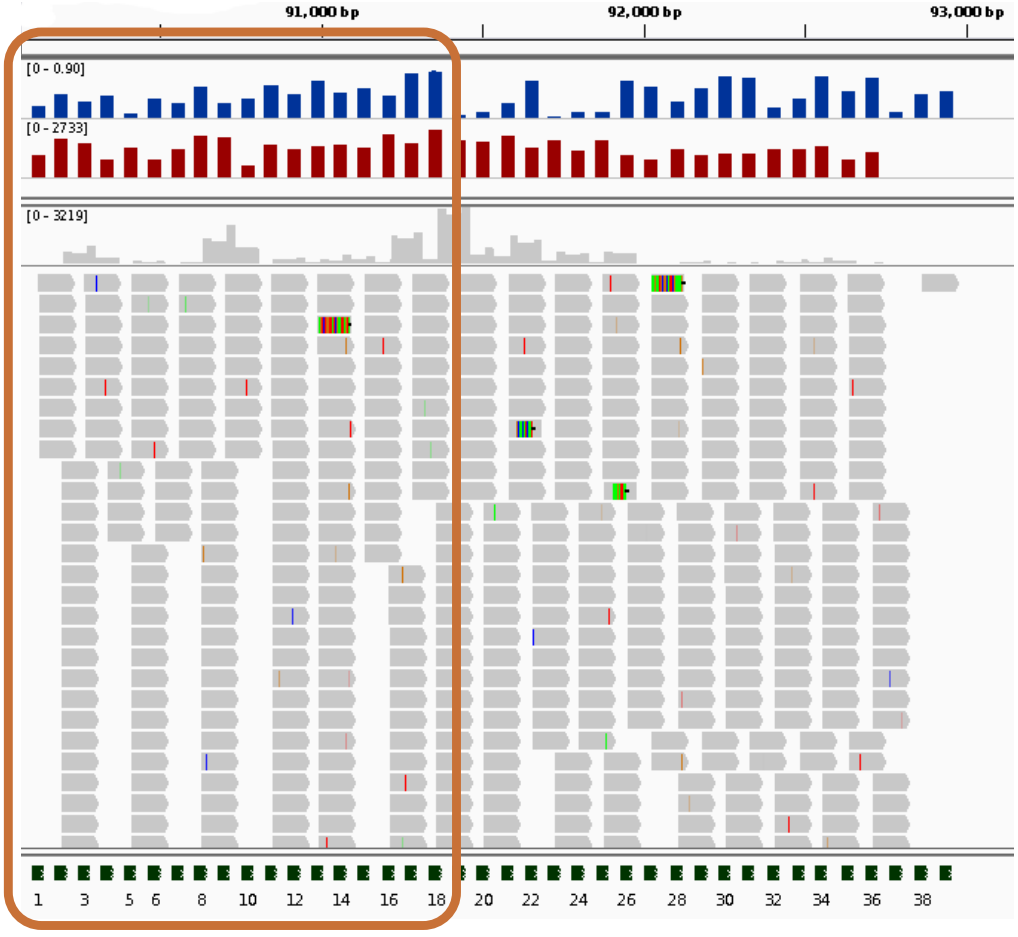
CRISPR3 of Synechocystis sp. PCC 6803



Spearman's correlation coefficient 0.58, p-value=0.011

Solution 1: Different Protocol

CRISPR3 of Synechocystis sp. PCC 6803

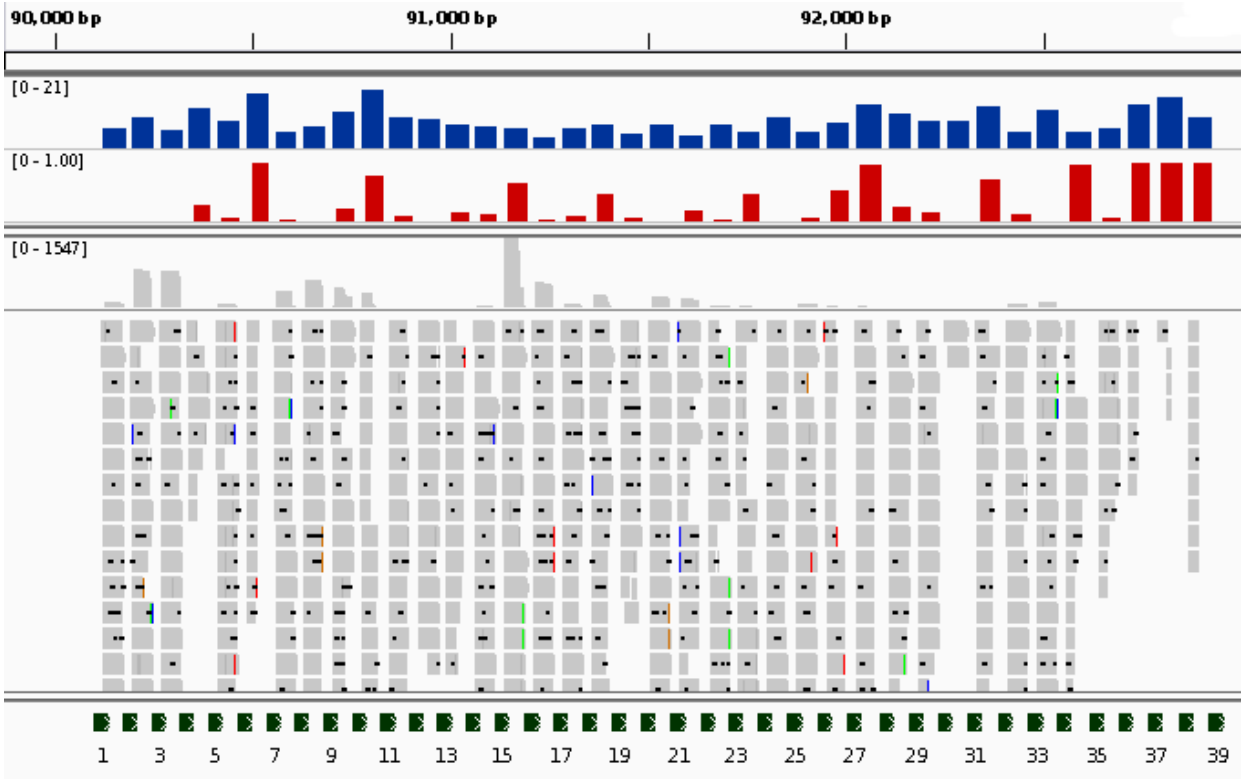


Spearman's correlation coefficient 0.58, p-value=0.011

- why only correlation in first eighteen?

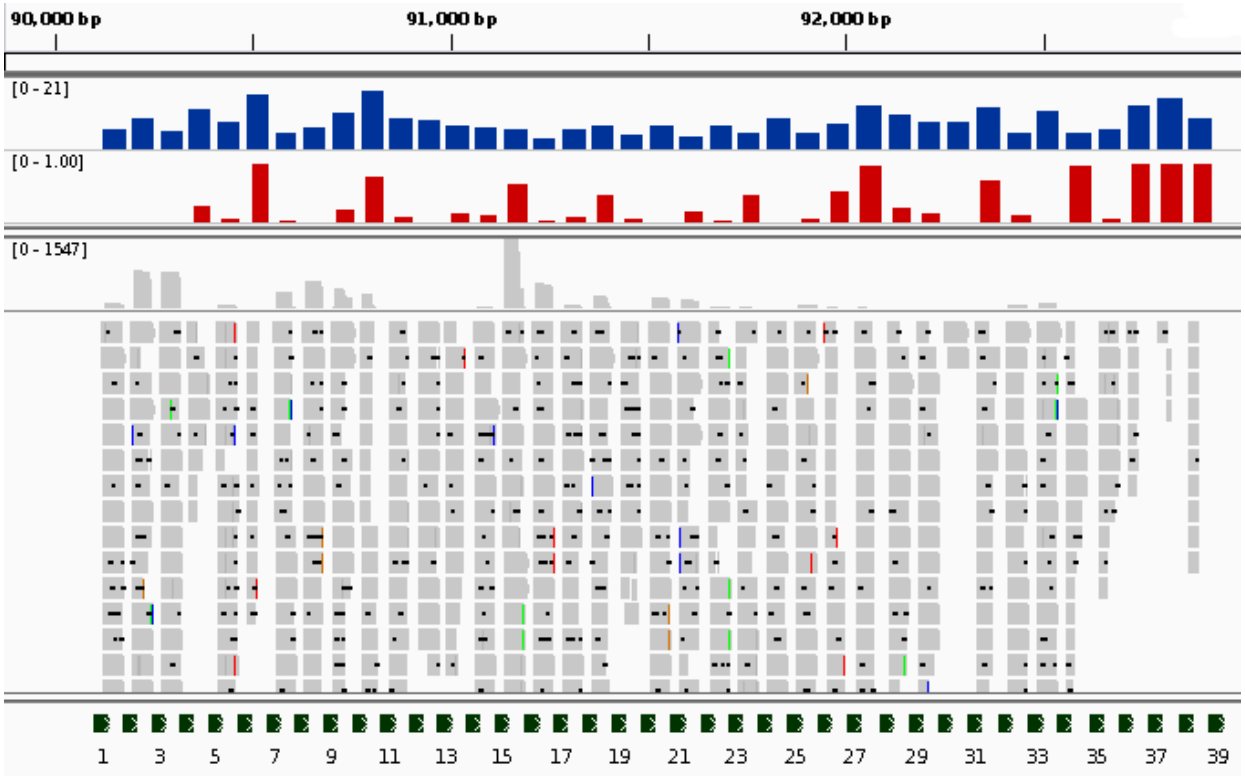
Solution 2: Different Question

- observation found read pattern for processed spacers
  - some always full length, others only part ⇒ **degradation?**
  - hence: look only on reads that cover at most one spacer
  - **question:** what is the fraction of full length/degraded spacers



Solution 2: Different Question

- observation found read pattern for processed spacers
  - some always full length, others only part ⇒ **degradation?**
  - hence: look only on reads that cover at most one spacer
  - **question:** what is the fraction of full length/degraded spacers

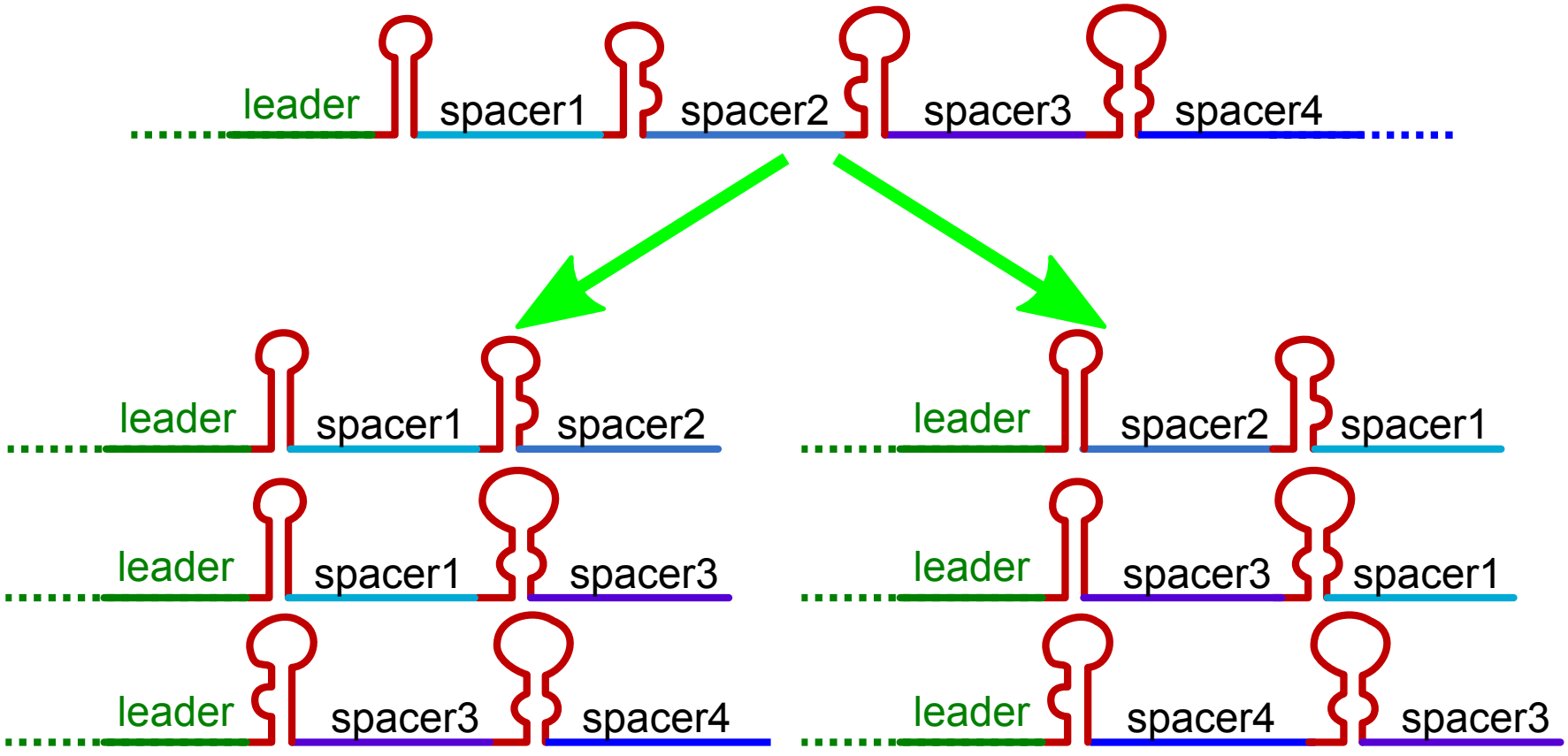


- **found:** high correlation to *structured-ness* of spacer

  stuctured-ness = low overall ensemble energy
- Pearson's correlation coefficient r=0.56 (p-value=0.00025).

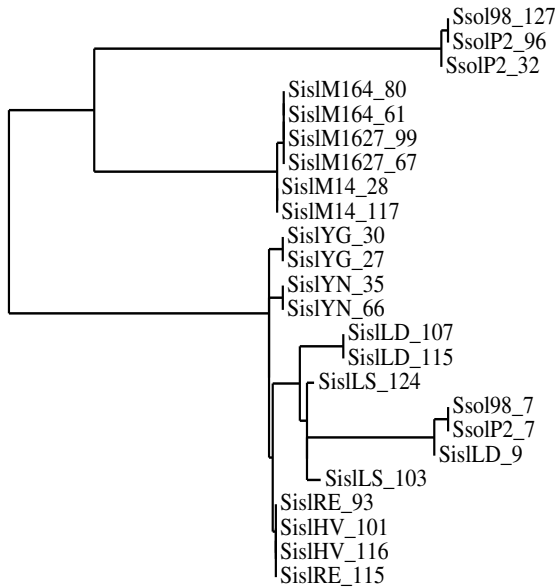
Solution 3: Combinatorial Library

- idea: get rid of position effects and precursor problems
- hence: look always at the first two spacers



CRISPR Leader

- 24 leader sequences (Crenarchaeal CRISPR - S.solfataricus and S.islandicus strains) with 200 nt (Shah et al.,2010)[1]
- Cluster based on sequence similarity (Blastclust)

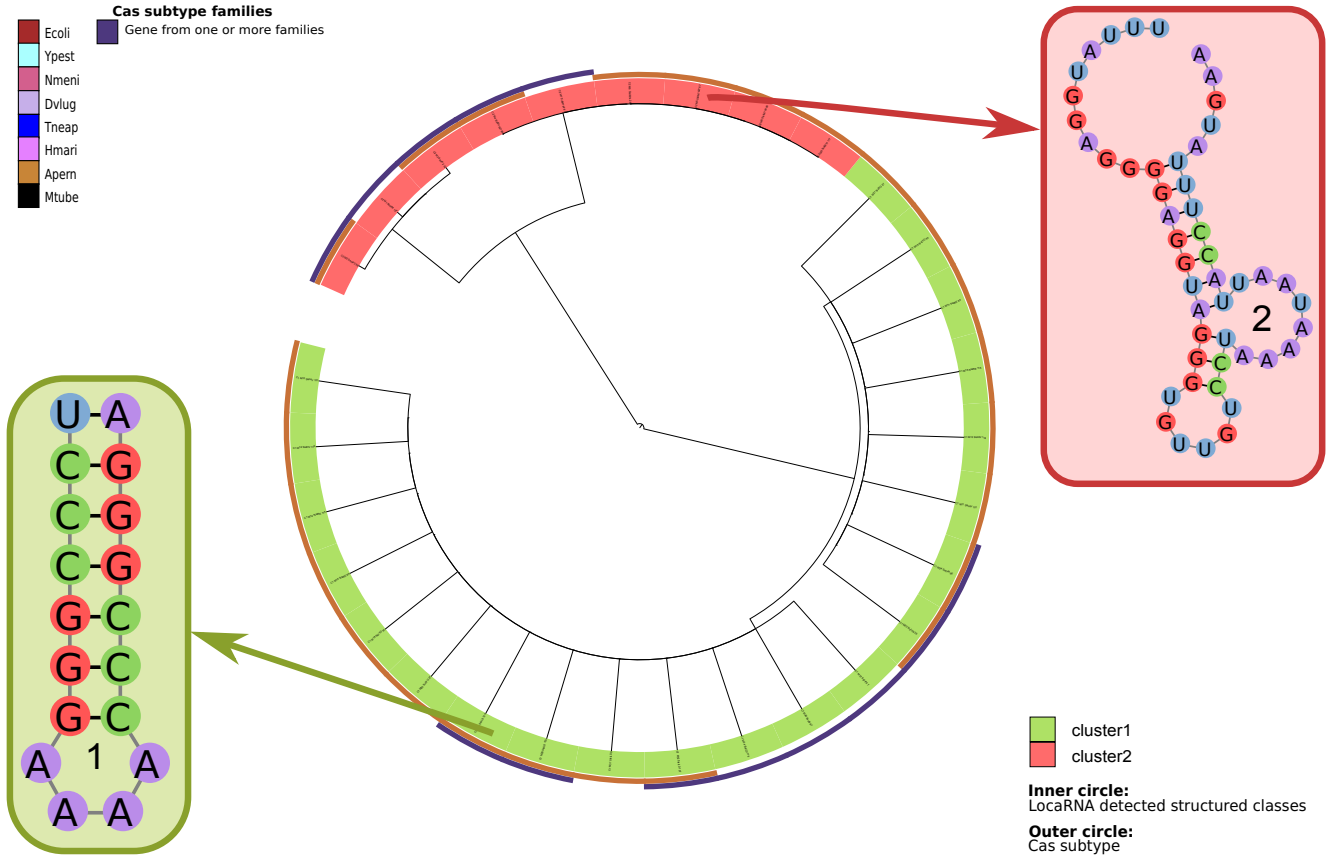


- Goal
  - Looking for secondary structure motifs.

[1] Shah SA, Garrett RA. CRISPR/Cas and Cmr modules, mobility and evolution of adaptive immune systems,Res Microbiol. 2011 Jan;162(1):27-38. Epub 2010 Sep 21

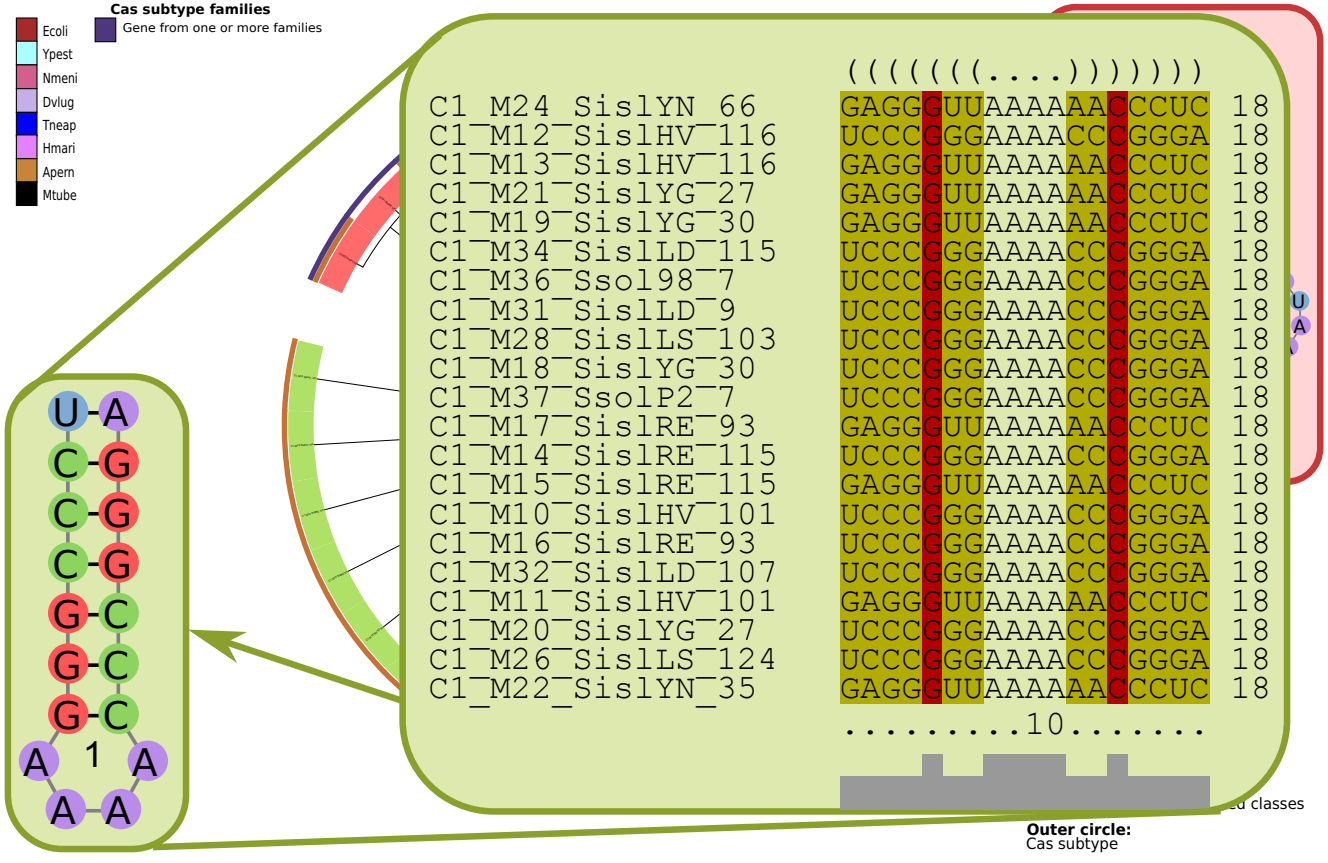
CRISPR Leader

- RNAMOTID: accuracy-based detection of local RNA elements
- 30 structure motifs are found.
- Cluster based on sequence structure similarity (LocaRNA).



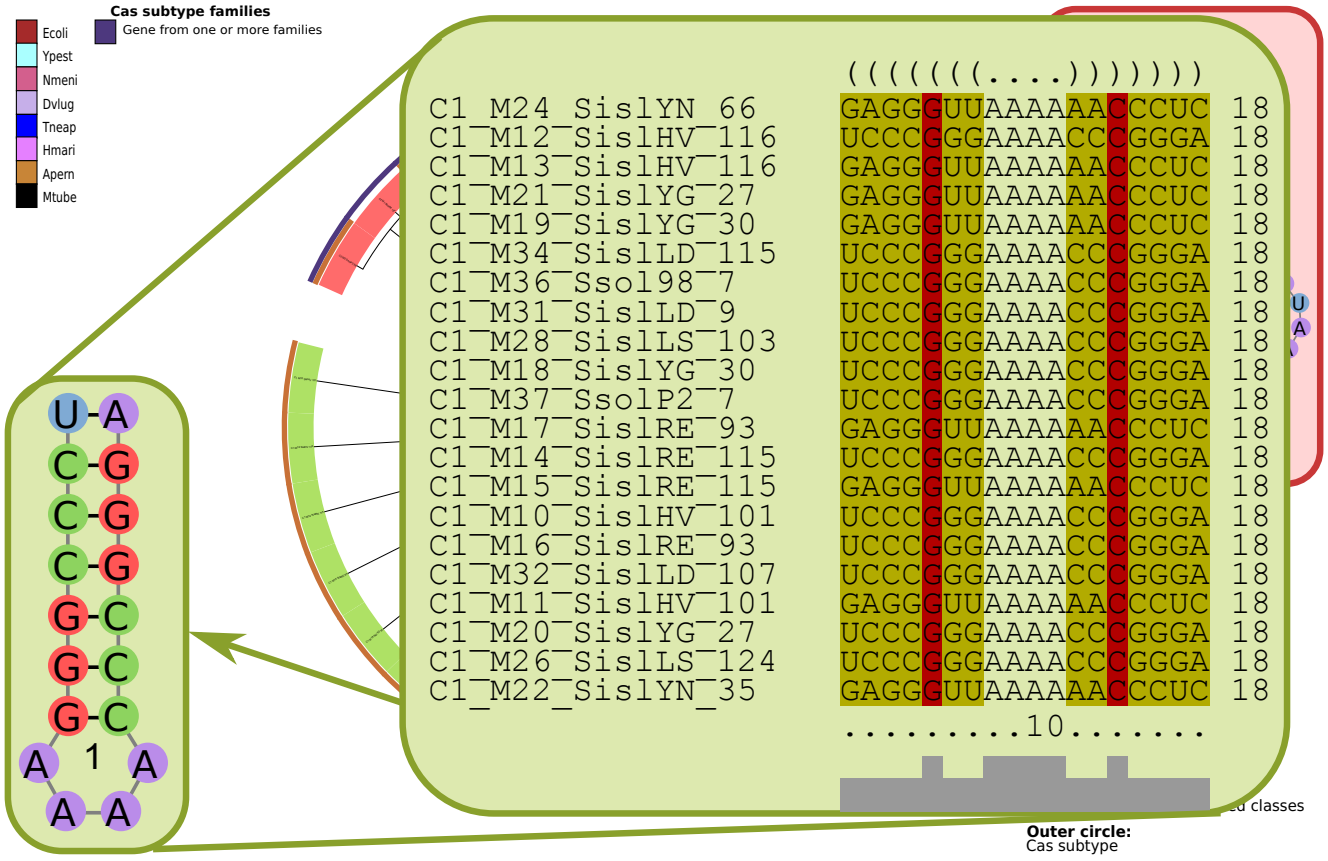
CRISPR Leader

- RNAMOTID: accuracy-based detection of local RNA elements
- 30 structure motifs are found.
- Cluster based on sequence structure similarity (LocaRNA).



CRISPR Leader

- RNAMOTID: accuracy-based detection of local RNA elements
- 30 structure motifs are found.
- Cluster based on sequence structure similarity (LocaRNA).



- **validation:** via covariance model using INFERNAL.
- it is specific for leader sequences

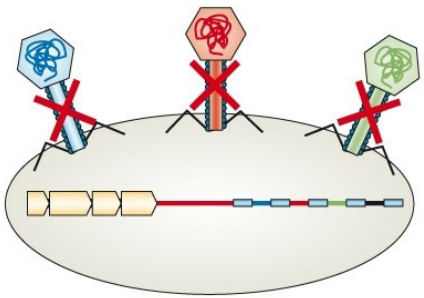
Conclusion

- structural clustering: evolutionary conserved structure
- structure quality and correlation of spacer reads
- surprise: correlation of (putative) degradation and spacer structure
- structural motif in leader sequences

Acknowledgements

- Sita Lange
- Omer S. Alkhnbashi
- Dominic Rose

- funding: Forschergruppe FOR 1680



Thank You for Your Attention