CNEFinder: Finding Conserved Non-coding Elements in Genomes

Lorraine A.K. Ayad[1] Solon P. Pissis[1]
Dimitris Polychronopoulos[2]

[1] King's College London

[2] Genomics England

Conserved Non-coding Elements

Conserved.
CNEs can be extremely conserved across evolution.

Conserved Non-coding Elements

Conserved.

CNEs can be extremely conserved across evolution.

Non-coding.

They do not encode for proteins. They are largely overlapping, with their genesis, functions and evolutionary dynamics being largely unknown.

Conserved Non-coding Elements

Conserved.
CNEs can be extremely conserved across evolution.

Non-coding.
They do not encode for proteins. They are largely overlapping, with their genesis, functions and evolutionary dynamics being largely unknown.

Elements.

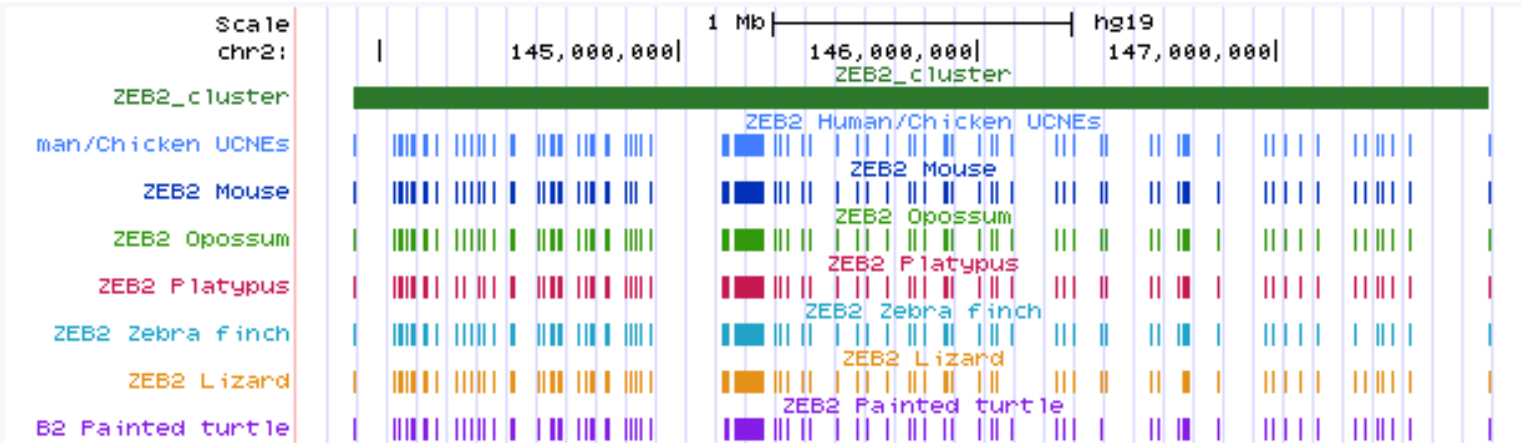


Figure: ZEB2 CNEs (UCNEBase)

Computing CNEs

Alignment Based methods - inspect pairwise or multiple whole genome alignments e.g BLASTZ , LASTZ

Computing CNEs

Alignment Based methods - inspect pairwise or multiple whole genome alignments e.g BLASTZ , LASTZ

Alignment Free methods - Warnefors et al. (2016) proposed an alignment-free method for the identification of CNEs in Drosophila based on k-mers.

k-mers occurring in the reference genome are mapped to the species of interest. Then overlapping hits are merged into longer CNEs.

Computing CNEs

Alignment Based methods - inspect pairwise or multiple whole genome alignments e.g BLASTZ , LASTZ

Alignment Free methods - Warnefors et al. (2016) proposed an alignment-free method for the identification of CNEs in Drosophila based on k-mers.

k-mers occurring in the reference genome are mapped to the species of interest. Then overlapping hits are merged into longer CNEs.

CNE Databases - contain already pre-computed sets of CNEs e.g Ancora, UCNEBase, cneViewer

Our Contribution

We present CNEFinder, a tool for identifying CNEs between two given DNA sequences with user-defined criteria.

Our Contribution

We present CNEFinder, a tool for identifying CNEs between two given DNA sequences with user-defined criteria.

It *does not* require or compute the whole-genome alignment of the two sequences.

Our Contribution

We present CNEFinder, a tool for identifying CNEs between two given DNA sequences with user-defined criteria.

It *does not* require or compute the whole-genome alignment of the two sequences.

It *does not* require or compute a whole-genome index such as the suffix array or the BWT.

CNEFinder

1. Identify seeds

CNEFinder

1. Identify seeds
2. Merge these seeds to form matches

CNEFinder

1. Identify seeds
2. Merge these seeds to form matches
3. Extend matches

1. Identify Seeds

Seeds of length $\lfloor \ell/(\ell - t \times \ell + 1) \rfloor$ are computed.
ℓ is the minimum CNE length.
$t \in (0, 1]$ is the relative identity threshold between the CNE pair.

1. Identify Seeds

Seeds of length $\lfloor \ell/(\ell - t \times \ell + 1) \rfloor$ are computed.
ℓ is the minimum CNE length.
$t \in (0, 1]$ is the relative identity threshold between the CNE pair.

This ensures each pair of elements with minimum length ℓ can have an edit distance of at most $\ell - t \times \ell$.

1. Identify Seeds

Seeds of length $\lfloor \ell/(\ell - t \times \ell + 1) \rfloor$ are computed.
ℓ is the minimum CNE length.
$t \in (0, 1]$ is the relative identity threshold between the CNE pair.

This ensures each pair of elements with minimum length ℓ can have an edit distance of at most $\ell - t \times \ell$.

Example

Let $\ell = 30, t = 0.7$

```
...ATTACAGCTAATTCAAACACTGCGGCGGTTGCTAT...
...TCCACTAAGCAACTTCAAACATGTCGCAGTTTCTCC...
```

2. Merge Seeds

The seeds found are then merged to produce co-linear sequences of non-overlapping matches.

2. Merge Seeds

The seeds found are then merged to produce co-linear sequences of non-overlapping matches.

The merging process is terminated once the addition of another gap would force the relative identity score to drop below threshold t.

2. Merge Seeds

The seeds found are then merged to produce co-linear sequences of non-overlapping matches.

The merging process is terminated once the addition of another gap would force the relative identity score to drop below threshold t.

Example

Let $\ell = 30, t = 0.7$

```
...ATTACAGCTAATTCAAACACTGCGGCGGTTGCTAT...
...TCCACTAAGCAACTTCAAACATGTCGCAGTTTCTCC...
```

3. Extend Matches

The current match is extended in both directions if the threshold allows it or otherwise in the direction having the smallest edit distance.

3. Extend Matches

The current match is extended in both directions if the threshold allows it or otherwise in the direction having the smallest edit distance.

This procedure is repeated until the computed relative identity score reaches threshold t or when the maximum length u of one of the elements has been reached.

3. Extend Matches

The current match is extended in both directions if the threshold allows it or otherwise in the direction having the smallest edit distance.

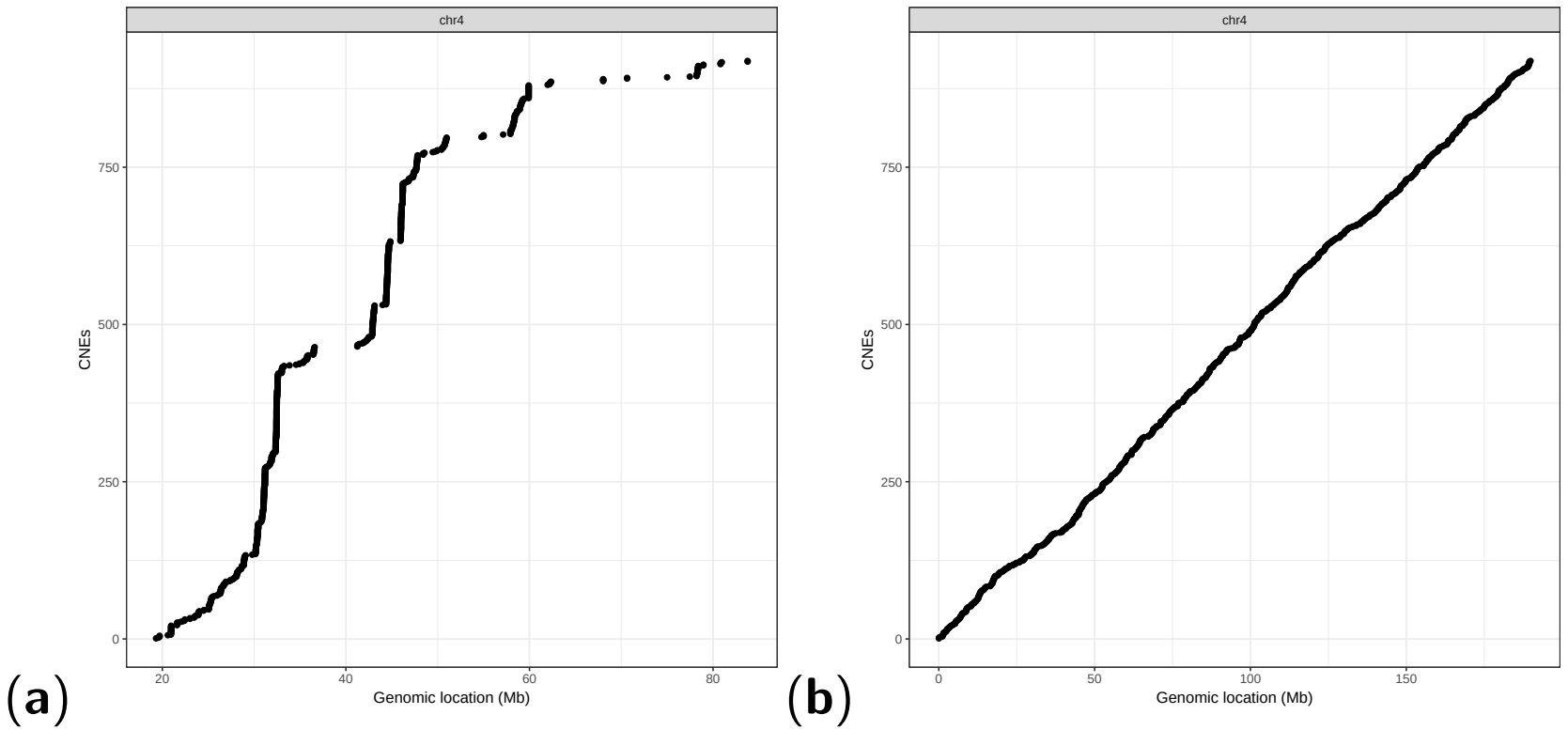
This procedure is repeated until the computed relative identity score reaches threshold t or when the maximum length u of one of the elements has been reached.

Example

Let $\ell = 30, t = 0.7$

```
...ATTACAGCTAATTCAAACACTGCGGCGGTTGCTAT...
...TCCACTAAGCAACTTCAAACATGTCGCAGTTTCTCC...
```

Genomic Distribution of CNEs



Genomic distribution of CNEs along human (hg38) Chr 4. (**a**) Elements found by CNEFinder. (**b**) CNE-like elements.

Results - Accuracy

	$200-250$ bp		$250-300$ bp		$300-350$ bp	
Gene	# CNEs Overlapping	% Nucleotides Overlapping	# CNEs Overlapping	% Nucleotides Overlapping	# CNEs Overlapping	% Nucleotides Overlapping
ZEB2	$31/31$	84.59	$18/18$	87.31	$20/20$	90.36
TSHZ3	$35/36$	78.49	$16/17$	80.01	$8/8$	84.85
EBF3	$28/28$	87.90	$17/17$	90.81	$16/16$	88.21
BCL11A	$20/20$	81.24	$28/28$	85.75	$14/14$	93.61
ZFHX4	$18/18$	88.02	$22/22$	89.82	$10/10$	86.86

Table: CNEs identified for five genes for different length ranges and $t = 0.95$.

Results - Accuracy

	$350-400$ bp		$400-450$ bp		$450-500$ bp	
Gene	# CNEs Overlapping	% Nucleotides Overlapping	# CNEs Overlapping	% Nucleotides Overlapping	# CNEs Overlapping	% Nucleotides Overlapping
ZEB2	$14/14$	83.17	$19/19$	91.56	$5/5$	92.45
TSHZ3	$6/6$	88.50	$12/12$	89.36	$2/2$	90.68
EBF3	$6/6$	78.46	$8/8$	83.91	$3/3$	82.21
BCL11A	$10/10$	90.73	$4/4$	83.49	$5/5$	88.04
ZFHX4	$6/6$	93.58	$5/5$	93.10	$6/6$	87.98

Table: CNEs identified for five genes for different length ranges and $t = 0.95$.

Results - Efficiency

- $143 - 148$ Mbp region of Chr 2 of the Human (hg19) genome
- $34 - 39$ Mbp region of Chr 7 of the Chicken (galGal3) genome
- 8 CPU cores and $t = 0.9$.

Results - Efficiency

- $143 - 148$ Mbp region of Chr 2 of the Human (hg19) genome
- $34 - 39$ Mbp region of Chr 7 of the Chicken (galGal3) genome
- 8 CPU cores and $t = 0.9$.

Length Range (bp)	200-250	250-300	300-350	350-400	400-450	450-500
Time (s)	4.4	4.4	4.5	4.8	4.3	5.2

Maximum memory used was 1.6 GB of RAM.

Results - Efficiency

- $143 - 148$ Mbp region of Chr 2 of the Human (hg19) genome
- $34 - 39$ Mbp region of Chr 7 of the Chicken (galGal3) genome
- 8 CPU cores and $t = 0.9$.

Length Range (bp)	200-250	250-300	300-350	350-400	400-450	450-500
Time (s)	4.4	4.4	4.5	4.8	4.3	5.2

Maximum memory used was 1.6 GB of RAM.

- $200 - 500$ bp with $t = 0.9$ and 8 CPU cores
- Whole Chr 2 of the human (hg19) genome
- Whole Chr 7 of the chicken (galGal3) genome

32m30s. Maximum memory 5.6 GB of RAM.

Comparison with Local Alignment Tools

We compared CNEFinder to YASS, local alignment search tool.

Comparison with Local Alignment Tools

We compared CNEFinder to YASS, local alignment search tool.

YASS works by identifying seeds between a pair of DNA sequences and then extends these matches to local alignments between the sequence pair.

Comparison with Local Alignment Tools

We compared CNEFinder to YASS, local alignment search tool.

YASS works by identifying seeds between a pair of DNA sequences and then extends these matches to local alignments between the sequence pair.

We ran YASS using:

- $76.57 - 79.01$ Mbp of Ch 8 of the Human (hg19) genome
- $123.57 - 124.82$ Mbp of Chr 2 of the Chicken (galGal3) genome.
- A dissimilarity threshold of 5%.
- These are the exact regions used to compute the CNEs for gene ZFHX4.

Comparison with Local Alignment Tools

We compared CNEFinder to YASS, local alignment search tool.

YASS works by identifying seeds between a pair of DNA sequences and then extends these matches to local alignments between the sequence pair.

We ran YASS using:

- $76.57 - 79.01$ Mbp of Ch 8 of the Human (hg19) genome
- $123.57 - 124.82$ Mbp of Chr 2 of the Chicken (galGal3) genome.
- A dissimilarity threshold of 5%.
- These are the exact regions used to compute the CNEs for gene ZFHX4.

For the elements identified by YASS, the average percentage of overlapping nucleotides was only **31.01%**.

Availability

https://github.com/lorrainea/CNEFinder

Publication

L. A. K. Ayad, S. P. Pissis, D. Polychronopoulos; CNEFinder: finding conserved non-coding elements in genomes, Bioinformatics, Volume 34, Issue 17, 1 September 2018, Pages i743-i747.