Non-Local Filtering of Alignment Seeds

Matthis Ebel[1]*, Giovanna Migliorelli and Mario Stanke*[1]

UNIVERSITÄT GREIFSWALD
Wissen lockt. Seit 1456

[1] Institute of Mathematics and Computer Science, University of Greifswald {matthis.ebel, mario.stanke}@uni-greifswald.de

Introduction

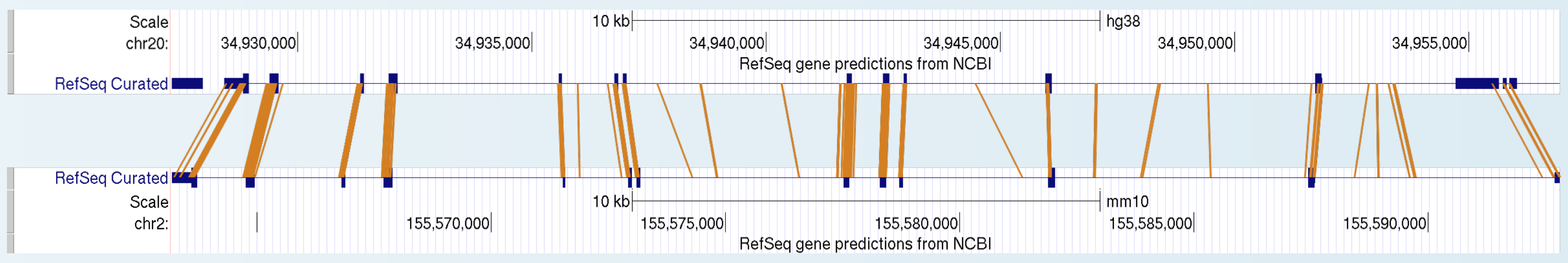
We present *geometric hashing* [2], a seed filtering method that quickly finds sensitive and specific sets of alignment seeds for two or more genomes.

A common approach for (multiple) genome alignment: seed and extend

- seed: seek short identical or similar fragments of the input sequences
- extend: create alignments starting from the seeds

Specific seeds with few false positives reduce the number of time consuming extensions and thus the overall runtime, while limiting the sensitivity (no seed, no alignment).

Geometric hashing finds clusters of seeds that may be thousands of basepairs apart but belong to the same (set of) ortholog genes. It is thus a suitable early step for de novo comparative genome annotation, e.g. of clades of related genomes.



Filtered seeds between human (top) and mouse (bottom) genes *glutathione synthetase*. Seeds (orange) mostly hit exons (blue) of the ortholog genes.

Edited screenshots from UCSC genome browser [4, 1]

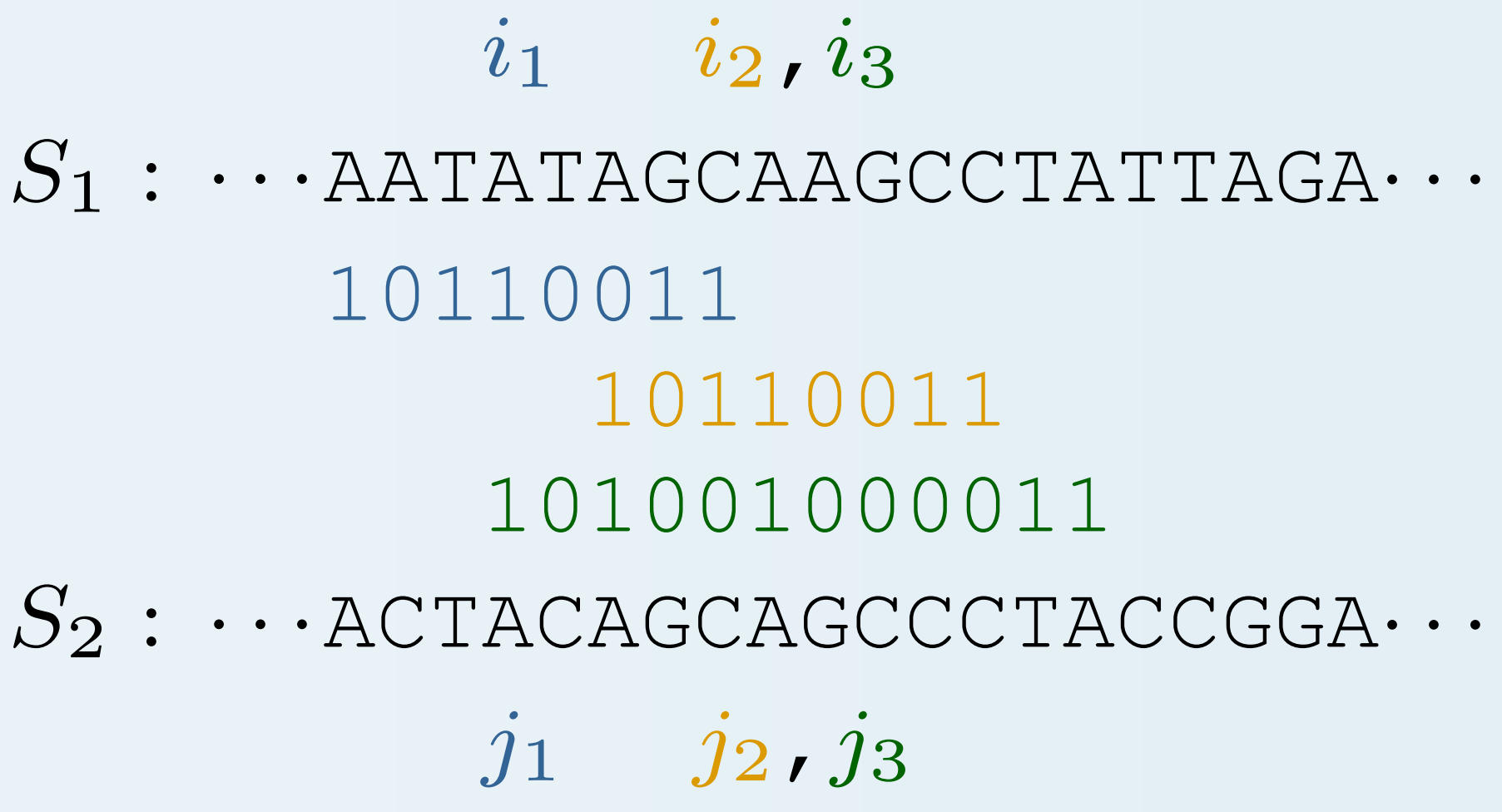
I – Seed Candidates

Seed candidates for geometric hashing can be simple exact k-mer matches of the input sequences, or spaced seeds that allow mismatches at fixed positions, possibly using multiple spaced seed patterns.

Our geometric hashing tool can be connected to Metagraph from Karasikov et al. [3], which allows efficient storage and fast querying of multiple sequences for exact shared k-mers and their coordinates. Our tool can use these k-mers to create spaced seeds.

Set of spaced seed patterns, $k = 5$: p_1 : `10110011`
p_1 : `101001000011`

`1` - “match position”, `0` - “don’t care position”

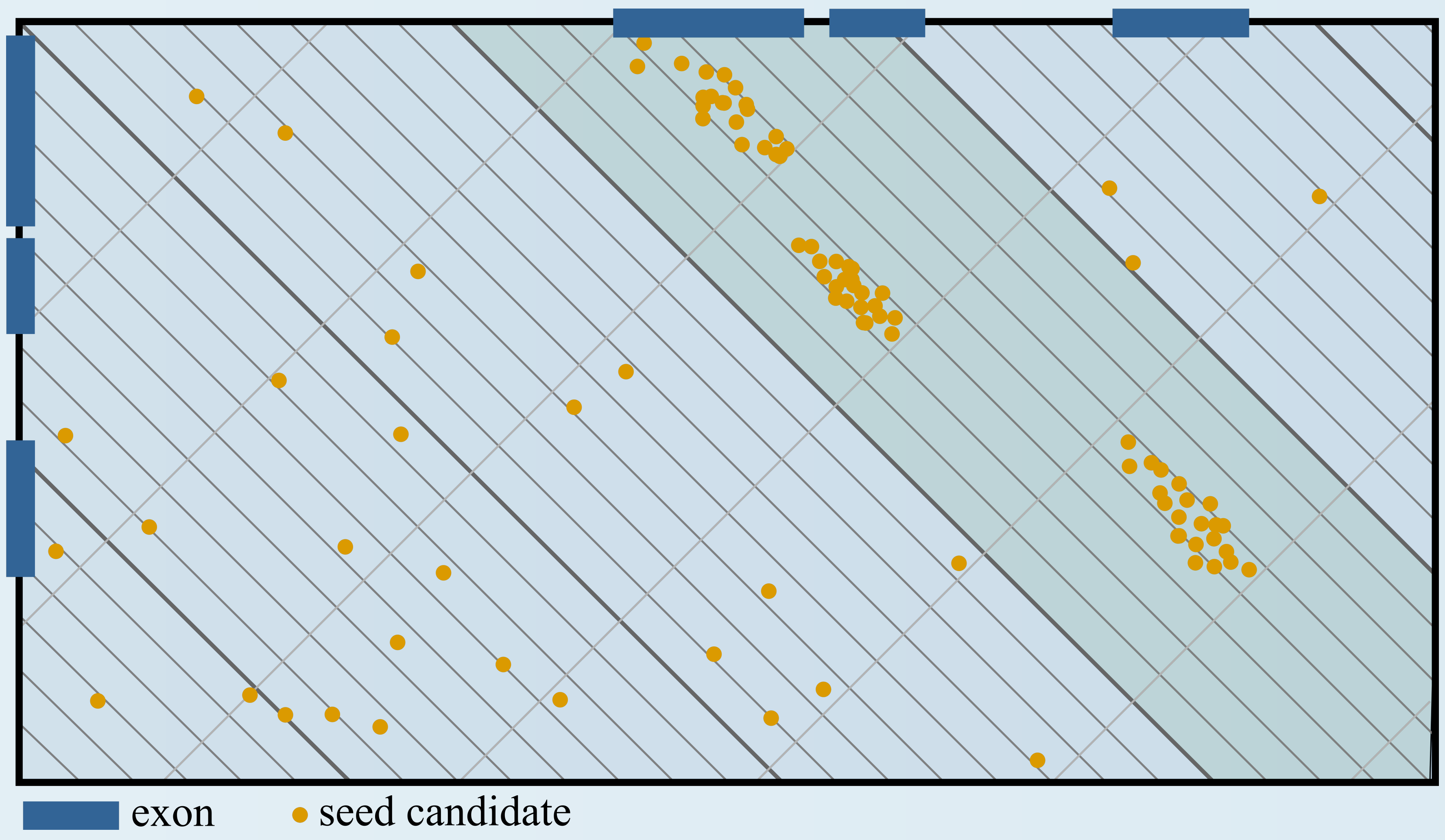

i_1 i_2, i_3
S_1 : ···AATATAGCAAGCCTATTAGA···
10110011
10110011
101001000011
S_2 : ···ACTACAGCAGCCCTACCGGA···
j_1 j_2, j_3

match	seed candidate
ATAGC	(S_1, i_1, S_2, j_1)
ACACC	(S_1, i_2, S_2, j_2)
AAATA	(S_1, i_3, S_2, j_3)
···	···

II – Geometric Hashing

Geometric hashing is a technique to find recurring patterns in data that may have undergone affine transformations such as relocation [5]. With it we efficiently identify sets of seed candidates that all have similar relative distances with respect to a reference genome. We use rounded relative distances to allow certain length variations of interjacent unconserved regions. Although we show here the special case of two input genomes, geometric hashing is easily extensible to multiple input genomes. All seed candidates that have a similar relative offset in their occurences appear in the same *tile* (big diagonal stripe with green shade in the right figure).

Tiles are divided into *chunks* (small diagonal boxes) and scored, where seed clustering in single chunks is beneficial: conserved exons of orthologous genes are expected to have many seeds, as opposed to an even distribution of false seeds that appear by chance. Only seeds from a high enough scoring tile are reported by geometric hashing, which greatly reduces false seeds compared to unfiltered seeds while keeping a high sensitivity.



Geometric hashing illustration in 2-dimensional alignment space.

$$Score(tile) = \frac{1}{\lambda^*} \sqrt[p]{\sum_{\text{chunks in tile}} n^p},$$

where n is the number of seed candidates in the respective chunk (small diagonal box), p is a parameter controlling the influence of seed candidate clusters and λ^* is a normalization term that is asymptotically proportional to the expected number of seed candidates if they were all evenly distributed.

III - Results

We selected a set of 705 orthologous genes from human and mouse, using flanked gene sequences of each gene as input. For each gene sequence, we added a random sequence of the same length to assess false positive seeds. Using optimized sets of spaced seeds patterns, we collected all seed candidates in our input data.

	k (4 patterns)	sensitivity	#FP	$\widehat{\text{FP}}$	add. runtime	memory
unfiltered	15	0.954	13,035,210	12	–	122 GB
geometric hashing	15	0.954	0	0	5 min	137 GB

$$\text{sensitivity} := \frac{\text{number of supported human CDS}}{\text{number of human CDS}} \qquad \widehat{\text{FP}} := \#\text{FP}\frac{N}{n_1 n_2} \quad \text{(normalized false positive count)}$$

A human CDS counts as “supported” if there is at least one seed that hits the CDS and also hits the orthologous mouse gene. #FP is the total number of FP seeds, i.e. seeds that match between two random sequences. n_1, n_2 are the total lengths of human and mouse input sequences, respectively, and N is the size of the human genome

References

[1] Jonathan Casper et al. “The UCSC genome browser database: 2018 update”. In: *Nucleic acids research* 46.D1 (2017), pp. D762–D769.

[2] Matthis Ebel, Giovanna Migliorelli, and Mario Stanke. “Global, highly specific and fast filtering of alignment seeds”. In: *BMC Bioinformatics* (2022).

[3] Mikhail Karasikov et al. “MetaGraph: Indexing and Analysing Nucleotide Archives at Petabase-scale”. In: *bioRxiv* (2020).

[4] W James Kent et al. “The human genome browser at UCSC”. In: *Genome research* 12.6 (2002), pp. 996–1006.

[5] Haim J Wolfson and Isidore Rigoutsos. “Geometric hashing: An overview”. In: *IEEE computational science and engineering* 4.4 (1997), pp. 10–21.

Read the paper:

https://rdcu.be/cPpXm