Alignment-free sequence comparison

Alignment-free sequence analysis

Although alignment-based approaches are most accurate and powerful for sequence comparison when they are feasible, their applications are limited in some situations.

First, for whole-genome comparison, there are many **duplications**, **translocations**, **large insertions/deletions**, and **horizontal gene transfers** (HGTs) in the genomes.

Second, in the current next-generation sequencing (**NGS**) era, investigators can sequence the genomes using NGS efficiently and economically. However, some parts of the genomes may not be sequenced due to the stochastic distribution of the reads along the genomes and the **difficulties of sequencing some parts of the genomes**, especially when the coverage is relatively low. Even if we can assemble the reads into long contigs, **these contigs may not share long homologous regions**, making it challenging to study the relationships among the genomes using alignment in such situations.

Third, **non-coding regions such as gene regulatory regions are not highly conserved** except for some functional regions, such as transcription binding sites, and cannot be reliably aligned.

Ren et al., 2018, Annual Review of Biomedical Data Science.

Alignment-free sequence comparison

Fourth, alignment is not suitable to compare sequences of large divergence.

Fifth, many large genome and metagenome data sets from shotgun NGS sequencing are available, and alignment-based methods are **too time consuming**.

Sixth, alignment-free methods are also **resistant to shuffling and recombination events** and are applicable when low sequence conservation cannot be handled reliably by alignment.

Seventh, the number of possible alignments of two sequences grows rapidly with the length of the sequences. **For two sequences of length N**, there are **$(2\ N)!/(N!)^2$ different gapped alignments**, which results in **about 10^{60} alignments for two sequences of length 100**).

Zielezinski et al., 2017, Genome Biology.

What is alignment-free sequence comparison?

Alignment-free approaches to sequence comparison can be defined as **any method of quantifying sequence similarity/ dissimilarity that does not use or produce alignment** (assignment of residue–residue correspondence) at any step of algorithm application.

Alignment-free approaches can be broadly **divided into two groups**:
(1) Methods based on the **frequencies of subsequences of a defined length (word-based methods)** and
(2) Methods that **evaluate the informational content between full-length sequences (information-theory-based methods)**.

There are also methods that cannot be classified in either of the groups, including those based on the **length of matching words (common, longest common, or the minimal absent words between sequences)**, **chaos game representation**, **iterated maps**, as well as **graphical representation of DNA sequences**, which capture the essence of the base composition and distribution of the sequences in a quantitative manner.

All of the alignment-free approaches are **mathematically well founded in the fields of linear algebra, information theory, and statistical mechanics**, and **calculate pairwise measures of dissimilarity or distance between sequences**.

Conveniently, most of these measures can be directly used as an input into **standard tree-building software, such as Phylip or MEGA**.

Zielezinski et al., 2017, Genome Biology.

How do word frequency-based methods work?

The rationale behind these methods is simple: **similar sequences share similar words/k-mers** (subsequences of length k), and mathematical operations with the words' occurrences **give a good relative measure of sequence dissimilarity**.

These methods first **count the number of occurrences of word patterns** (*k*-mers, *k*-grams, *k*-tuples) along a sequence.

Secondly, a **similarity/dissimilarity measure** is defined between any pair of sequences based on the word count frequencies.

Finally, various **clustering algorithms** such as hierarchical clustering and neighbor joining are used to **group the sequences**.

In practice, the **word size (k) of 2–6 residues** produces stable and optimal protein sequence comparisons across a wide range of different phylogenetic distances.

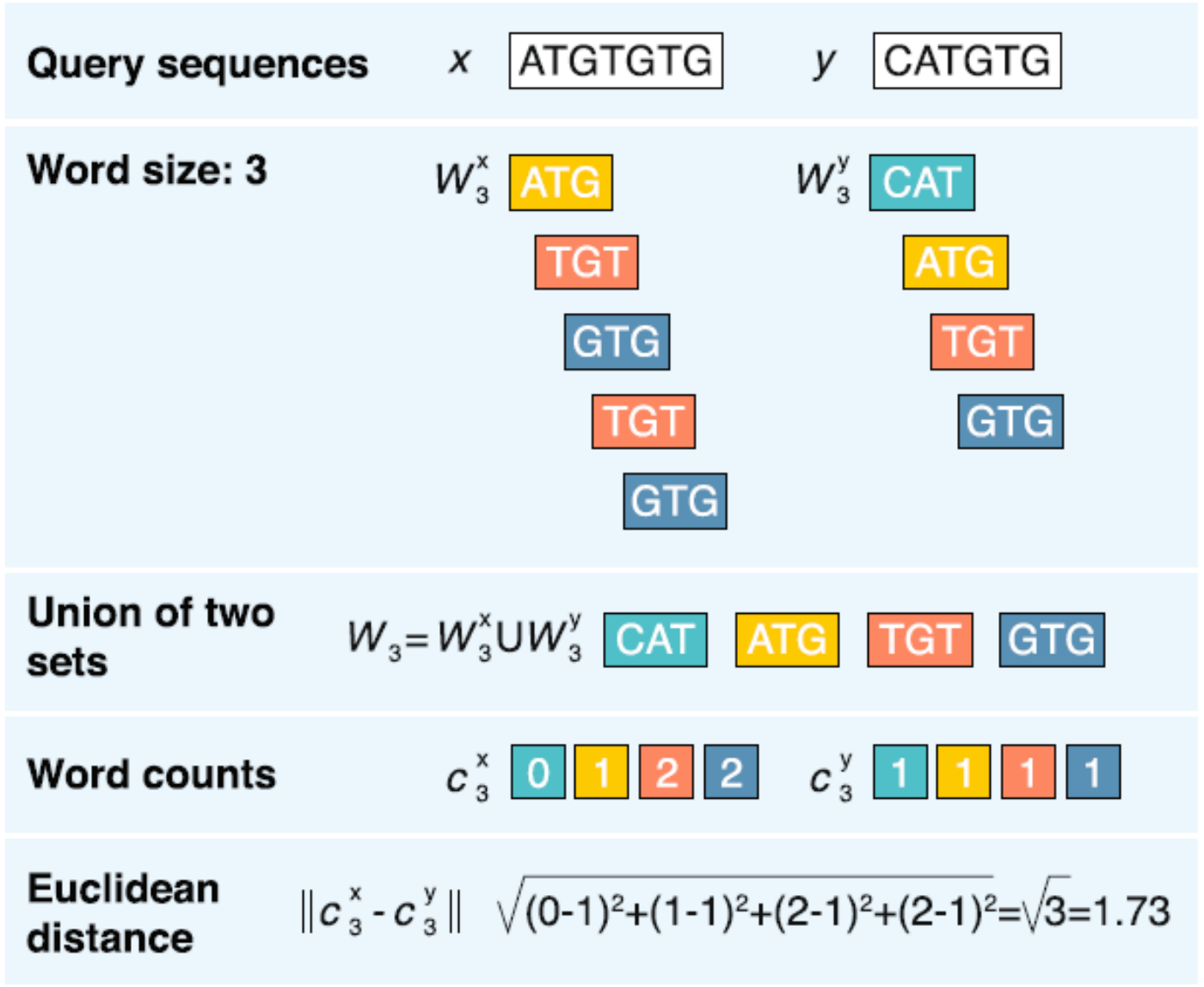
In nucleotide sequence analyses, **k can safely be set to 8–10 for genes or RNA, 9–14 bases for general phylogenetic analyses**.

Up to 25 bases in case of comparison of isolates of the same bacterial species.

As a rule of thumb, **smaller k-mers should be used when sequences are obviously different** (e.g., they are not related) whereas **longer k-mers can be used for very similar sequences**.

Ren et al., 2018, Annual Review of Biomedical Data Science.

How do word frequency-based methods work?



Zielezinski et al., 2017, Genome Biology.

How do word frequency-based methods work?

Alternatively, **DNA/RNA or protein alphabet can be reduced to a smaller number of symbols based on chemical equivalences**. This procedure may increase the detection of homologous sequences that display very low identity. For example, the **four-letter DNA alphabet can be distilled to two-letter purine–pyrimidine encoding**, and **proteins can be represented by 5, 4, 3, or even 2 letters according to their different physical–chemical properties**.

The second step (**mapping sequences onto vectors**) is by far the most customizable; **instead of using vectors of word counts or word frequencies**, there are many other ways to create vectors, **which range from weighting techniques to normalization and clustering**.

Additionally, because word-based methods operate on vectors, their mathematical elegance allows the employment of more than 40 functions other than the Euclidean distance, such as the Pearson correlation coefficient, Manhattan distance, Google distance, etc.

Zielezinski et al., 2017, Genome Biology.

Similarity/ dissimilarity measurement: examples

Pearson correlation coefficient

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where,

r = Pearson Correlation Coefficient

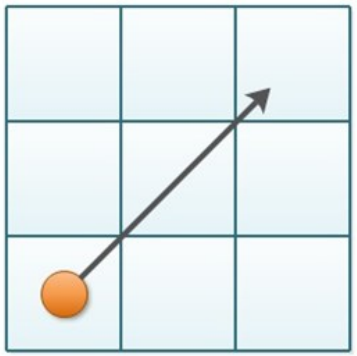
x_i = x variable samples

y_i = y variable sample

$\bar{x}$ = mean of values in x variable

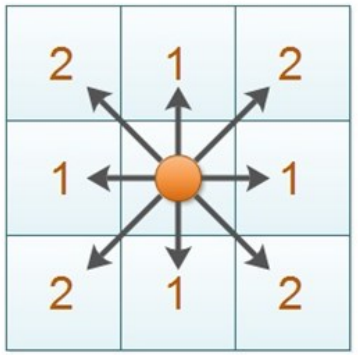
$\bar{y}$ = mean of values in y variable

Euclidean Distance



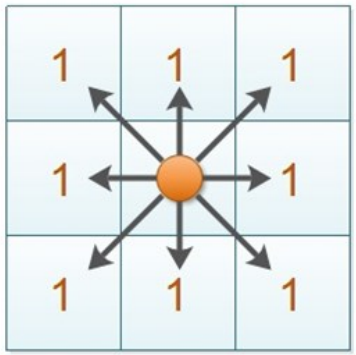
$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Manhattan Distance



$$|x_1 - x_2| + |y_1 - y_2|$$

Chebyshev Distance



$$\max(|x_1 - x_2|, |y_1 - y_2|)$$

Zielezinski et al., 2017, Genome Biology.

How do information theory-based methods work?

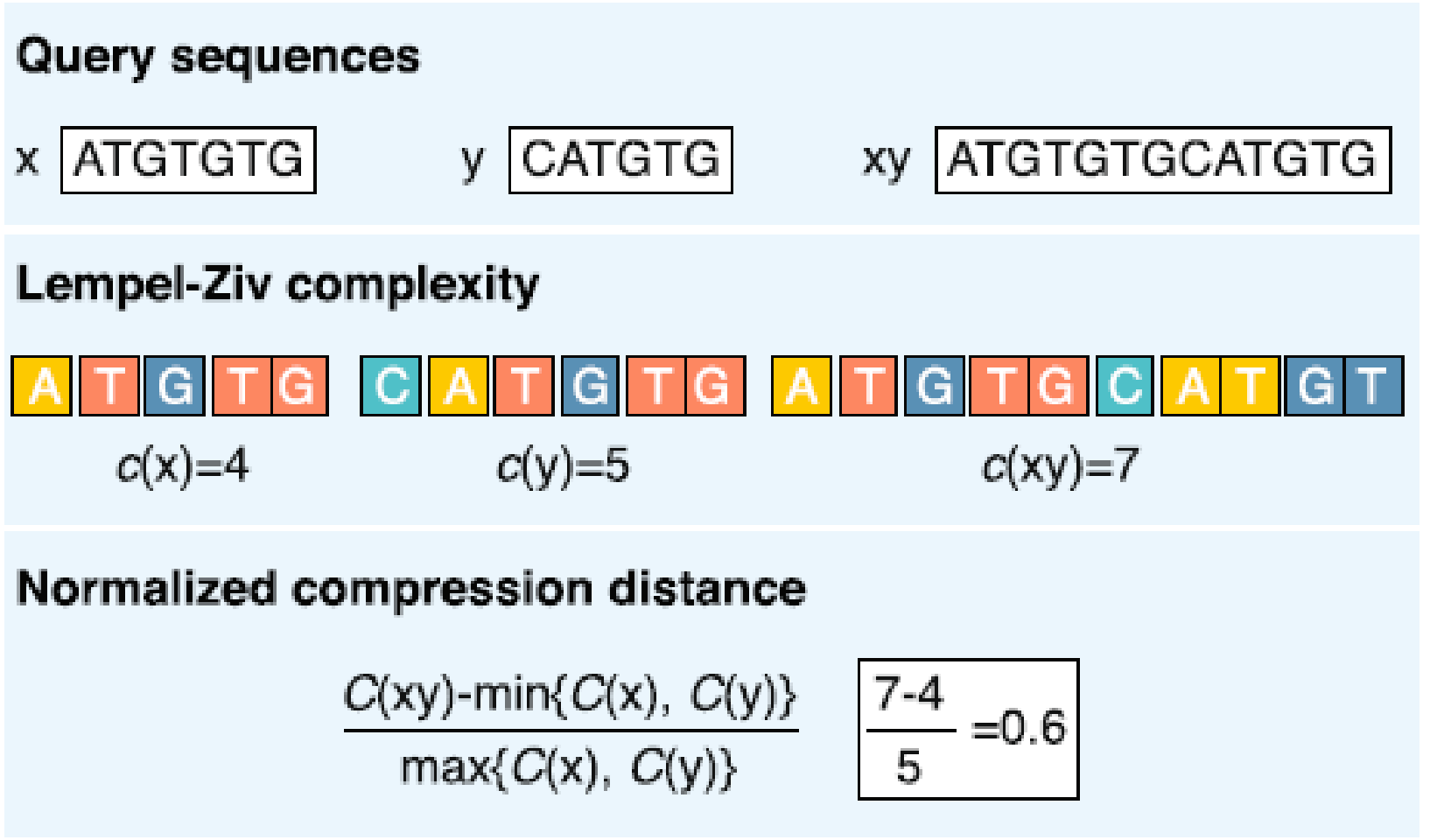
Information theory-based methods **recognize and compute the amount of information shared between two analyzed biological sequences**. Nucleotide and amino acid sequences are ultimately strings of symbols, and their digital organization is naturally interpretable with information theory tools, such as **complexity and entropy**.

For example, the **Kolmogorov complexity** of a sequence can be measured by the **length of its shortest description**. Accordingly, the sequence **AAAAAAAAAA** can be described in a few words (**10 repetitions of A**), whereas **CGTGATGT presumably has no simpler description than specification nucleotide by nucleotide (1 C, then 1 G and so on). Intuitively, longer sequence descriptions indicate more complexity**.

However, Kolmogorov did not address the method to find the shortest description of a given string of characters. Therefore, the complexity is most commonly approximated by general compression algorithms (e.g., as implemented in .zip or .gzip programs) where the length of a compressed sequence gives an estimate of its complexity (i.e., a more complex string will be less compressible).

Zielezinski et al., 2017, Genome Biology.

How do information theory-based methods work?

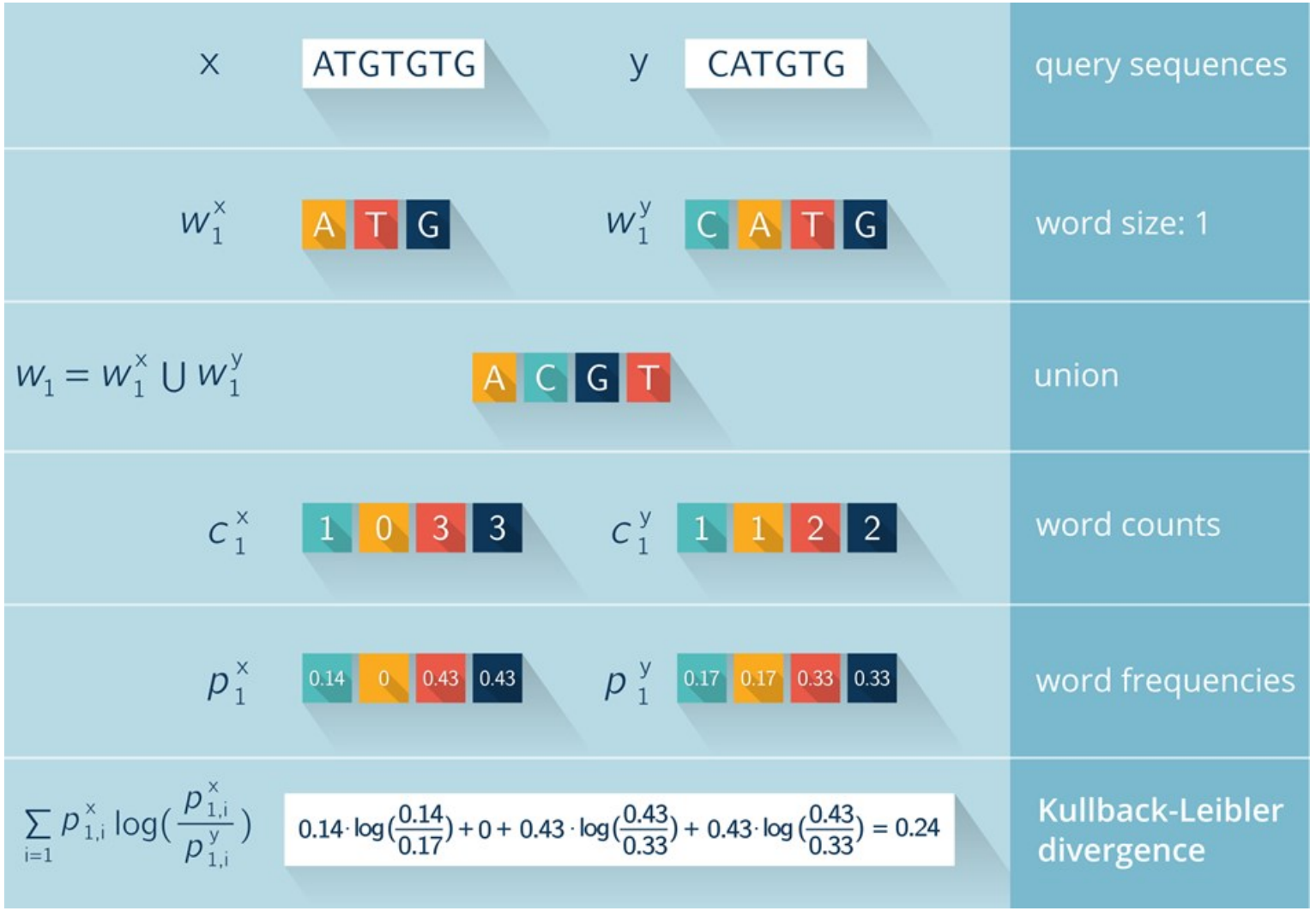


Alignment-free calculation of the normalized compression distance using the Lempel–Ziv complexity estimation algorithm. Lempel–Ziv complexity counts the number of different words in sequence when scanned from left to right (e.g., for s = ATGTGTG, Lempel–Ziv complexity is 4: A|T|G|TG).

Zielezinski et al., 2017, Genome Biology.

How do information theory-based methods work?

Another example of an information measurement often applied to biological sequences is entropy.



Zielezinski et al., 2017, Genome Biology.

How are alignment-free methods used in next generation sequencing data analysis?

The data volume of samples sequenced so far (**estimated to be only 10^{-20}% of the total DNA on Earth**) is already challenging the storage and processing capacities of modern computers.

These tools build an index of k-mers from a reference set of transcripts and then calculate the expression by matching them to each sequencing read directly. **Such "pseudoalignment" describes the relationship between a read and a set of compatible transcripts**. Grouping pseudoalignments belonging to the same set of transcripts allows one to directly infer the expression of each transcript model. This approach to quantify gene/ transcript expression levels from **RNA sequencing reads is both 10–100 times faster than any of the alignment-based methods** and at least as accurate as best performing alignment-based workflows.

Zielezinski et al., 2017, Genome Biology.

Where else can alignment-free sequence comparison methods be applied?

Distantly related, **remote sequences that evolve beyond recognizable similarity are one of the most classic applications of alignment-free mastering**.

Another rising trend for the use of word-based alignment-free methods is the **detection of functional and/or evolutionary similarities among regulatory sequences (e.g., promoters, enhancers, and silencers) to estimate their *in vivo* activities in different organisms (flies and mammals, including humans)**.

Alignment-free measures were also applied to **detect domain shuffling signatures in proteins and to identify the members of complex multidomain proteins, such as kinases**.

Horizontal gene transfer strongly complicates the task of reconstructing the evolutionary history of genes and species, and alignment-free methods have also proved to be helpful in this field.

Whole-genome phylogeny is another area where alignment-free methods play an increasing role.

Sequence classification is another field that might benefit from bringing together different alignment-free approaches, such as **grouping expressed sequences tags** that originate from the same locus or gene family, **clustering expressed sequence tag sequences with full-length cDNA data**, and aggregating gene and protein sequences into functional families.

Zielezinski et al., 2017, Genome Biology.

How well do alignment-free methods work?

In general, tested alignment-free methods can be **as good as alignment algorithms** and may **perform even better in case of protein sequences that underwent domain shuffling events**.

Zielezinski et al., 2017, Genome Biology.

Thank You