

Supplementary Table 1 – Comparison of SEQSIM with representative alignment-free (AF) and pairwise similarity methods. Methods are evaluated based on their algorithmic approach, emphasis on contiguous sequence matches, output format, typical use cases, and their relation to SEQSIM. This comparison highlights SEQSIM's unique ability to detect long, high homologous regions in promoter sequences at genome scale.

Method	Alignment-Free Approach	Contiguous Match Emphasis	Output	Typical Use Cases	Relation to SEQSIM
SEQSIM	Needleman-Wunsch inspired scoring. No dynamic programming; rewards long matches.	High – prioritizes extended high identity regions	Pairwise similarity index (%)	Promoter / promoter comparisons across genome	
N² (1)	k-mer frequencies with mismatch “neighborhoods” (word vector correlation)	Low – counts shared k-mers (mismatches allowed), not requiring contiguity	Similarity/distance matrix (% similarity)	Enhancer/promoter clustering; motif content comparison	Broad motif-level similarity; complements SEQSIM by finding functional similarity even without long exact matches. Code no longer available through publication.
ACS(2)	Longest exact common substrings (averaged length)	High – requires exact contiguous matches (no mismatches)	Distance or similarity (%) based on average match length	Closely related sequences; genomic phylogeny when moderate conservation exists	Extreme case of SEQSIM (no mismatches tolerated). Could validate if SEQSIM's similarities involve exact runs. Less useful if sequences have mutations.
KMACS (3)	Longest common substrings allowing <i>k</i> mismatches (k-mismatch ACS)	High – rewards long nearly contiguous regions (few mismatches allowed)	Distance matrix (alignment-free phylogenetic distance)	Sequence sets with some divergence; phylogeny of genes/genomes with mutations	Seems to be comparable to SEQSIM - both highlight long high-identity regions. Could not run a sample as source code is no longer available on http://kmacs.gobics.de/

FSWM (4)	Filtered spaced-word matches (pattern-based ungapped alignment)	High – finds local alignments under a spaced mask (tolerates mismatches at “don’t care” positions)	Distance matrix or dendrogram (phylogenetic)	Whole-genome or gene set comparisons; finds homologous regions with substitutions	Similar target as SEQSIM (local homology) with different technique. Good cross-check for SEQSIM-detected segments, especially with mismatches. No longer available on http://fswm.gobics.de/
Clustal Omega (5)	NOT Alignment Free – performs MSA using guide trees and HMM profile alignments	Medium–High – alignment-based; rewards global similarity with gaps allowed	Multiple sequence alignment and pairwise percent identity matrix	Protein and nucleotide MSAs; phylogenetics; sequence homology visualization	Not alignment-free, but its percent identity matrix offers a familiar comparison format. SEQSIM's similarity matrix shows visual and numerical similarity to Clustal Omega's output (Figure 4), justifying its use as a reference baseline. Of all mentioned methods, Clustal is the most widely documented and cited.
MatGAT (6)	Pairwise global alignments (PAM/BLOSUM scoring), no MSA required	Medium–High – scores extended similarity using scoring matrices	Similarity/identity matrix (%)	Small sets of DNA/protein sequences without MSA	Closest in spirit to SEQSIM; similar output and pairwise comparison, but not very scalable and not truly AF. (See Supplementary Figure 3)
Mash (7)	MinHash sketch of k-mers (Jaccard similarity of k-mer sets)	Medium – long exact matches yield many shared k-mers, but mismatches break k-mers	Distance or similarity (% identity estimate, with p-value)	Massive comparisons (10k+ sequences); fast clustering, database search	Fast screening tool; highlight similar overall patterns to SEQSIM but some major differences can be seen (See Supplementary Figure 3) due to difference in algorithm.

Supplementary Table References:

1. Göke J, Schulz MH, Lasserre J, Vingron M. Estimation of pairwise sequence similarity of mammalian enhancers with word neighbourhood counts. *Bioinformatics*. 2012 Mar 1;28(5):656–63.
2. Ulitsky I, Burstein D, Tuller T, Chor B. The average common substring approach to phylogenomic reconstruction. *J Comput Biol*. 2006 Mar;13(2):336–50.
3. Leimeister CA, Morgenstern B. Kmacs: the k-mismatch average common substring approach to alignment-free sequence comparison. *Bioinformatics*. 2014 Jul 15;30(14):2000–8.
4. Leimeister CA, Sohrabi-Jahromi S, Morgenstern B. Fast and accurate phylogeny reconstruction using filtered spaced-word matches. *Bioinformatics*. 2017 Apr 1;33(7):971–9.
5. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol*. 2011 Oct 11;7:539.
6. Campanella JJ, Bitincka L, Smalley J. MatGAT: An application that generates similarity/identity matrices using protein or DNA sequences. *BMC Bioinformatics*. 2003 Jul 10;4(1):29.
7. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*. 2016 Jun 20;17(1):132.