

# Maximum Likelihood Estimation of Weight Matrices for Targeted Homology Search

Peter Menzel<sup>1,2,\*</sup>, Jan Gorodkin<sup>1</sup>, and Peter F. Stadler<sup>2-6</sup>

<sup>1</sup>Division of Genetics and Bioinformatics, IBHV, University of Copenhagen, Grønnegårdsvej 3, DK-1870 Frederiksberg, Denmark

<sup>2</sup>Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany.

<sup>3</sup>Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany

<sup>4</sup>Fraunhofer Institut für Zelltherapie und Immunologie, Perlickstraße 1, D-04103 Leipzig, Germany

<sup>5</sup>Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Vienna, Austria

<sup>6</sup>The Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, New Mexico

\*Corresponding Author

**Abstract:** Genome annotation relies to a large extent on the recognition of homologs to already known genes. The starting point for such protocols is a collection of known sequences from one or more species, from which a model is constructed – either automatically or manually – that encodes the defining features of a single gene or a gene family. The quality of these models eventually determines the success rate of the homology search. We propose here a novel approach to model construction that not only captures the characteristic motifs of a gene, but are also adjusts the search pattern by including phylogenetic information. Computational tests demonstrate that this can lead to a substantial improvement of homology search models.

## Introduction

Homology search is one of the generic important tasks in bioinformatics. It is indispensable, e.g., for the assessment of the phylogenetic distribution of genes and gene families and it forms the basis for detailed phylogenetic analyses in general. Homology search also comprises the first step in gene annotation pipelines. The ever increasing influx of genomic sequence data makes reliable and automated homology search a crucial bottleneck in many projects.

Typically, the starting point for homology search is a collection of known sequences, usually in the form of a multiple sequence alignment. Then, one or all of these “seed sequences” are fed into a pairwise alignment algorithm – such as `blast` [MM04] – and compared to the sequence database of the target species. In many cases, e.g. for distant

homologs or short query sequences, the sensitivity of this approach is too low. In such cases one can determine from the alignment the sites that share the same residues in all or most of the seed sequences. These highly conserved sequence blocks typically comprise the specific biological function of the gene – like binding site motifs, catalytically active sites, or structural elements. Once identified, these blocks can be used to build a more sophisticated search pattern that contains the intrinsic properties of this particular gene. The *fragrep* approach, for instance, represents the query as a collection of short consensus patterns and distance constraints between them [MSS06]. Again, restricting oneself to the consensus sequence information of the blocks may lead to a rather low sensitivity or specificity of the search pattern. This is the case e.g. for DNA binding sites [Sto00], which not necessarily share a common consensus sequence.

More expressive sequence models can be build with position specific scoring matrices (PSSM), which record the relative frequencies of residues at each site. The application of PSSMs for homology search requires more elaborate profile alignment algorithms. An example for proteins is *psi-blast* [AMS<sup>+</sup>97]. For short, ungapped, PSSMs arising e.g. as models of transcription factor binding sites, a relative scoring scheme is used [KGR<sup>+</sup>03], which can be extended to the gapped case by means of fractional programming [MCS07]. Hidden Markov Models are a viable alternative. In many cases, the highly variable gap sizes and the small set of seed sequences are problematic for the training procedures. PSSM-based approaches therefore were instrumental in several recent studies on highly variable ncRNA families such as Y RNAs [MGSS07], vault RNAs [SCH<sup>+</sup>09], and telomerase RNAs [XMQ<sup>+</sup>08].

While theoretically straightforward, the construction of reliable PSSMs from sequence alignments turns out to be a quite non-trivial task. In principle, one just has to count the frequency of the residues in the alignment columns, decide on a scheme to treat gap characters, and possibly add pseudo-counts. In practice, however, one has to deal with biases in the phylogenetic distribution of the seed sequences, which are often dominated by a set of closely related model organisms. The small size of the seed set, on the other hand, makes it undesirable to exclude a large fraction of the available data. A commonly used remedy is to use one of several weighting schemes [VS93]. For amino acid sequences more sophisticated methods for creating unbiased PSSMs are available, e.g. via the *EasyPred* web server [Nie]. Such unbiased “centroid” PSSMs, however, still do not include all the available phylogenetic information, in particular, they do not take into account any knowledge on the relative phylogenetic position of the target genome among the aligned seed sequences.

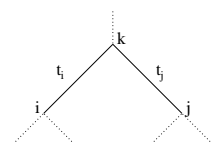
In this contribution we therefore explore the possibility to employ a maximum likelihood (ML) approach to optimize search patterns for usage on a particular target. Our approach is similar in spirit to the reconstruction of ancestral sequences from their extant offsprings. Given a phylogenetic tree  $T$ , ancestral genes are “resurrected” by inferring the states for internal nodes of  $T$  given the known sequences at the leafs. The earliest approaches were based on the parsimony principle [Fit71]. Alternatively, maximum likelihood methods, introduced by Felsenstein [Fel81], are in use. The latter require an explicit model of sequence evolution. On the other hand, they naturally provide probability distributions over the amino acid or nucleotide alphabet for every sequence position and every internal node

of the tree. In other words, ML provides us with PSSMs for ancestral states. Compared to parsimony approaches, maximum likelihood methods are more accurate because branch lengths, more detailed residue substitution models, and back-mutations are taken into account [ZJ97]. Ancestral sequence reconstruction has been proven to be a powerful tool for testing hypotheses regarding the function of genes from extinct species, see, e.g., [Tho04].

Here, we modify this approach. Instead of focusing on the internal nodes of the tree  $T$ , we use the same mathematical machinery to infer the most likely nucleotide sequence at an additional leaf node in the tree — the target species for homology search.

## Construction of Search Patterns

We start from a given multiple sequence alignment  $M$  with  $m$  sequences and a phylogenetic tree  $T$  with  $m + 1$  leaves, representing the phylogenetic relationships and branch lengths among the  $m$  species included in the alignment, and a single additional target species 0. Our approach combines two ML computations. First we use  $M$  and  $T \setminus 0$ , the phylogenetic tree restricted to the aligned species, to estimate for each alignment column  $i$  a relative substitution rate  $\hat{\mu}_i$ . The calculation of the likelihood follows Felsenstein’s pruning algorithm [Fel81]. The likelihood of a residue  $s_k$  at an interior node  $k$  is obtained from the corresponding likelihoods at the two child nodes  $i$  and  $j$ , which are separated from  $k$  by branches of length  $t_i$  and  $t_j$ , respectively:



$$L_{s_k}(\mu) = \left( \sum_{s_i} P_{s_k s_i}(t_i, \mu) L_{s_i}(\mu) \right) \times \left( \sum_{s_j} P_{s_k s_j}(t_j, \mu) L_{s_j}(\mu) \right) \quad (1)$$

For each alignment column  $i$ , we numerically optimize  $\hat{\mu}_i = \operatorname{argmax}_{\mu} L_T(\mu)$  using Golden Section Search [Kie53]. The likelihood of the tree  $T$  is given by the sum over all possible states  $s_r$  at the root node  $r$ :  $L_T(\mu) = \sum_{s_r} \pi_s L_{s_r}(\mu)$  where the  $\pi_s$  are the prior probabilities of observing letter  $s$ . The transition matrix  $\mathbf{P}$  contains probabilities  $P_{xy}(t, \mu) = [e^{t\mu\mathbf{Q}}]_{xy}$  for changing from state  $y$  to state  $x$  over time  $t$  and a rate  $\mu$ . The instantaneous rate matrix  $\mathbf{Q}$  represents a standard substitution model, such as the HKY85 [HKY85] or General Time Reversible (GTR) [Tav86] model for DNA sequences. Parameters for these models can be estimated from the alignment by using standard maximum likelihood analysis software like PAML [Yan07]. We advocate that this should be done ideally on larger data sets than the usually short query alignments themselves.

In the second step, we use the estimated values  $\hat{\mu}_i$  to compute the probabilities for each residue at the  $i$ -th position of the target sequence. To this end, we re-root the original tree  $T$  to the target species 0 and then calculate the likelihoods  $L_{s_0}(\hat{\mu}_i)$  for  $T^0$ . From these likelihoods at the root node of  $T^0$ , we directly obtain the residue probabilities in each alignment column  $i$ . Finally, these probabilities are transformed into a PSSM.

Figure 1 exemplifies the difference of a PSSM inferred by the ML approach and a PSSM obtained by counting the nucleotide frequencies in the seed alignment. In this particular

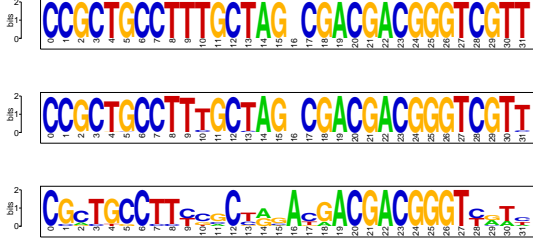


Figure 1: Example for estimating a PSSM. **top:** Target sequence in the 5' region of the 7SK RNA of *Drosophila persimilis*. **middle:** Maximum likelihood estimated nucleotide probabilities for *D. persimilis*. **bottom:** PSSM derived from nucleotide frequencies of the 11 other *Drosophila* sequences.

case, the ML estimate is significantly more informative and much closer to the motif in the target sequence.

The ML-PSSM pattern depends explicitly on the relative position of the target species in  $T$ . If the target is in close proximity to one or more other species, then high probabilities will be assigned to the residues that are present in those neighboring species. With increasing distance to the target species, on the other hand, the probabilities will converge to an uninformative equilibrium distribution. A column equilibrates faster, the larger the substitution rate  $\hat{\mu}_i$ . The algorithm thus tells us, which alignment columns or regions can be expected to be informative for a particular target sequence. To this end, we compute the Shannon information of each alignment position as

$$H(i) = - \sum_s f_i(s) \cdot \log_2 f_i(s) \quad (2)$$

where  $f_i(s)$  is the estimated frequency of residue  $s$  at position  $i$ . The corresponding information content is  $I(i) = \bar{H} - H(i)$ , where  $\bar{H} = - \sum_s \bar{f}(s) \log_2 \bar{f}(s)$  and  $\bar{f}(s)$  is the background distribution of the residues. In the simplest case,  $\bar{H} = 2$  for an uniform distribution of the four nucleotides.

Significant patterns can now be extracted by finding windows of a user-defined minimum length that have an average information content above a certain threshold. Alignment columns with high estimates of  $\hat{\mu}$ , on the other hand, can be excluded from the search pattern to compensate for highly variable sites. Thus, the maximum likelihood algorithm not only provides residue probabilities for each alignment position, but also gives information about the conserved sites and the variation of mutation rates within one sequence. We remark that our approach of optimizing the  $\hat{\mu}_i$  is similar to the method used in the Rate4Site program [PBM<sup>+</sup>02], which aims at identifying functional important regions in protein surfaces.

## Performance Evaluation

As test data we used a collection of genomic `multiz` alignments of *Drosophila* species [Con07] downloaded from the UCSC Genome Browser<sup>1</sup>. Only segments covering all 12

<sup>1</sup><http://hgdownload.cse.ucsc.edu/goldenPath/dm3/multiz15way/>

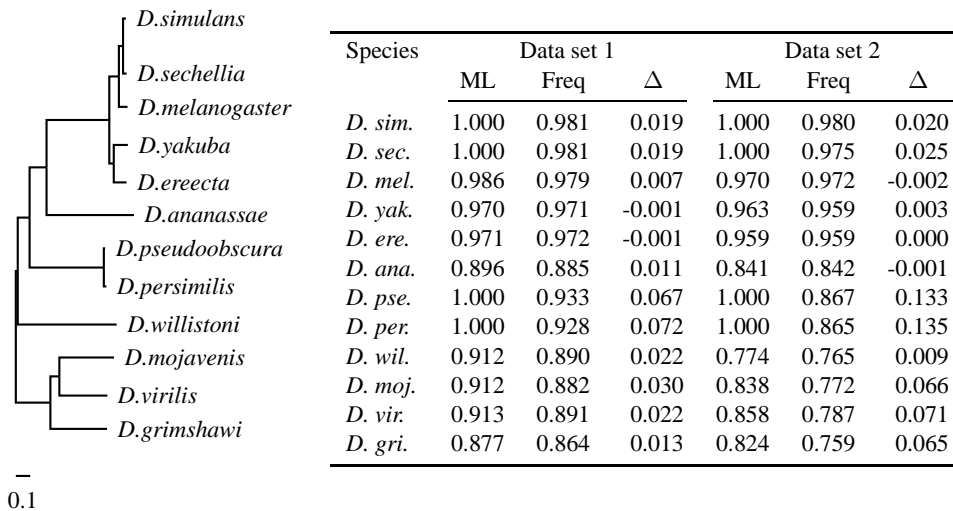


Figure 2: **left:** Phylogenetic tree of the 12 *Drosophila* species [Con07]. **right:** Median match scores of the maximum likelihood PSSMs (ML) and the frequency PSSMs (Freq) for 10 randomly selected 30nt windows from each alignment in both data sets.

drosophilid species were retained and gapped columns excluded. *Set1* consists of the 56 alignment segments of *D. melanogaster* chromosome 4 with minimum length 500 and a `multiz` score of at least 10000. The average pairwise sequence identity is 76.1%. *Set2* contains 45 alignments with `multiz` scores between 100 and 10000 and minimum length of 200. This set has 67.1% average sequence identity.

We removed one sequence at a time from the alignment and computed the residue probabilities for this sequence with our ML approach from the 11 remaining sequences using the phylogenetic tree in figure 2 and the HKY85 substitution model. The transition bias parameter  $\kappa$  was estimated using PAML. For comparison, we computed the position frequency matrix from the same 11 species. Both results were converted to a PSSM. From each alignment we randomly selected 10 windows of different lengths. The MATCH scores [KGR<sup>+</sup>03] of the corresponding interval of the two PSSMs against the 12th aligned sequence that was excluded from training were computed using `pwmatch`<sup>2</sup> [TBF<sup>+</sup>07]. Then we compared the match scores of each pair of PSSMs and used the Wilcoxon rank-sum test to see if the maximum likelihood (ML) scores are significantly larger than the scores from the frequency method (Freq).

Figures 3 and 4 show the MATCH scores of each pair of PSSMs for windows of length  $L = 30$  for *Set1* and *Set2* for a representative subset of the 12 drosophilid species. Overall, we observe that the ML matrices have significantly higher MATCH scores than the frequency matrices for most of the target species. The difference is especially apparent for those drosophilids that have a closely related neighbor in the phylogenetic tree, such as *D. simulans* and *D. sechellia* or *D. pseudoobscura* and *D. persimilis*. Here the median

<sup>2</sup><http://www.bioinf.uni-leipzig.de/Software/pwmatch/>

MATCH score improvement is up to 0.076 for *D. persimilis* in *Set1* and 0.135 in *Set2*. Only for *D. ananassae* and *D. willistoni* there is no significant difference of the scores in *Set2* where both the ML and Freq PSSMs perform equally and only a slight average improvement of the ML PSSMs is visible in *Set1*. Due to the relatively large distance from all other species, and the relatively even distribution of the species in the tree, the frequency-based matrix scores are very similar to the ML estimate in these two cases. Generally, the improvement of the MATCH scores is higher in *Set2*, which has lower sequence identity. For instance, the average score difference of both methods in *D. pseudoobscura* is 0.067 in *Set1* and 0.133 in *Set2*, where the median score of the frequency method is much lower than in *Set1*.

For homology search, short blocks with high information content are of particular importance, since such queries can be searched most efficiently. Thus, we extracted from both data sets those sub-patterns containing columns with high information content at most positions. Figure 5 summarizes the MATCH scores of the ML and the frequency PSSMs for all (non-overlapping) windows of length 20nt which have an average information content of at least 1.8 bits in the ML matrices. For these patterns, we observe again that the ML approach performs significantly better for most target species. For some species, only few windows fulfilling these constraints can be found, e.g. *D. ananassae* (n=23) or *D. willistoni* (n=27). Due to the relatively large distance to the other drosophilids, the ML algorithm assigns high residue probabilities only to highly conserved alignment columns. Eventually, these probabilities are very similar to the nucleotide frequencies in the seed alignment and the performance of ML and frequency approach becomes indistinguishable.

Due to the close phylogenetic relationship of *D. simulans* and *D. sechellia*, and *D. pseudoobscura* and *D. persimilis*, resp., the ML approach estimates very high nucleotide probabilities for these target species. Thus many windows with high average information content can be found. Compared to the frequency PSSMs, the ML PSSMs provide a big performance improvement in these species.

## Discussion

In this contribution we presented a novel approach for constructing PSSM-like sequence models for homology search. Unlike standard methods, our maximum likelihood method aims at building models that are specifically adapted to a particular target species. This is achieved by utilizing the phylogenetic information of the seed sequences and the relative position of the target species therein.

Evaluation on genomic sequence alignments of the 12 sequenced drosophilid species shows that the maximum likelihood method indeed provides the expected improvements. We are able to find highly conserved sites in the alignment and make use of the sequence information from neighboring species in the phylogenetic tree. The more proximal a known sequence is related to the target species, the more specific the search pattern from the maximum likelihood computation becomes, even for randomly drawn samples. If the target species is evolutionary distant in the tree from the known taxa, the alignment

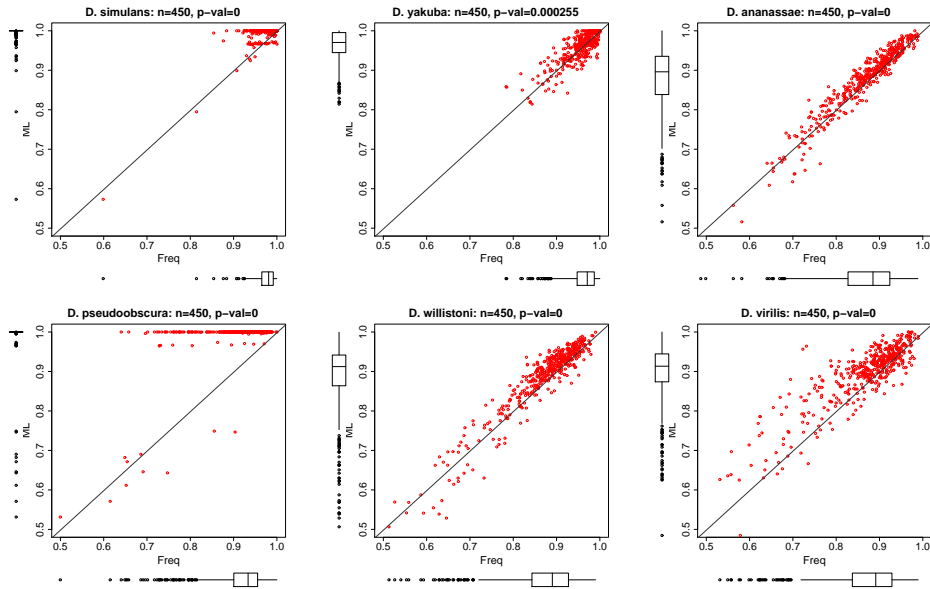


Figure 3: *Set 1*: MATCH scores of maximum likelihood (ML) and frequency (Freq) PSSMs for random windows of length 30nt ( $n = 450$ ). P-values of “0” are smaller than machine precision.

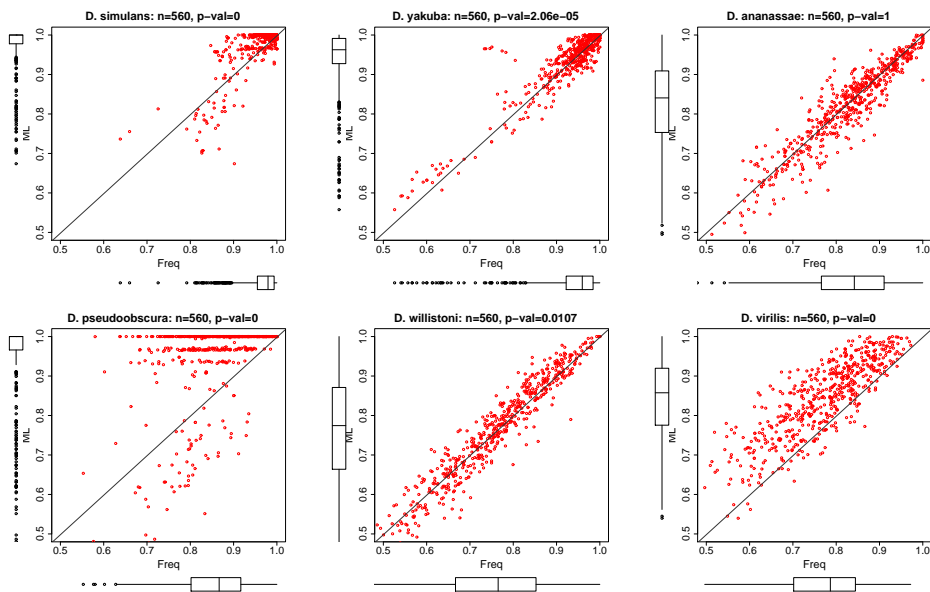


Figure 4: *Set 2*: MATCH scores for maximum likelihood (ML) and frequency (Freq) PSSMs for random windows of length 30nt ( $n = 450$ ). P-values of “0” are smaller than machine precision.

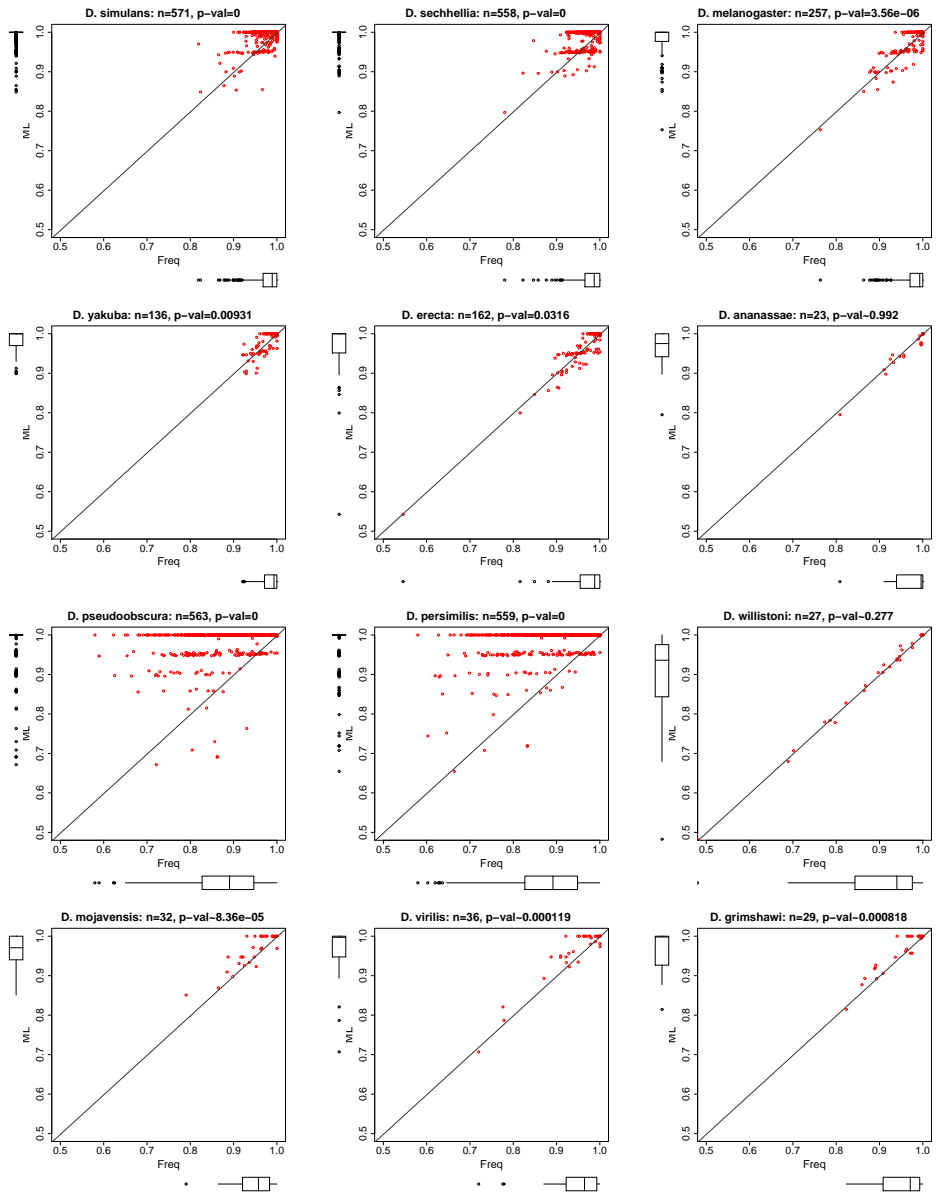


Figure 5: *Set2*: MATCH scores for the ML and frequency-based PSSMs for all non-overlapping windows of length 20 with average information content  $H \geq 1.8$ . In a few cases (indicated by  $p \sim$  instead of  $p =$ ) the  $p$ -value estimates are approximations due to the small sample size.



sites with high information content can be used for building the search pattern and the specificity is better or the same compared to normal search patterns based on residue frequencies.

The approach proposed here is potentially useful not only for the purely sequence-based homology search. In particular for structured RNAs it seems natural to incorporate phylogenetic information also into covariance models such as those utilized by SCFG-based tools. To this end, base pair substitution models for paired alignment columns need to be incorporated. We expect that this will be helpful in the detection of conserved structural elements in ncRNA families as well as aiding in automatic estimation of highly probable structure motifs in a target species. A second issue that needs to be addressed in future work is the handling of gaps, which we excluded here for the sake of clarity. In the simplest case, the approach of `fragrep` [MCS07] provides a remedy.

**Acknowledgement.** PM is supported by the Danish research council for Technology and Production through and the Danish research school in biotechnology. This work was supported by the Danish Center for Scientific Computation.

## References

- [AMS<sup>+</sup>97] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25:3389–3402, 1997.
- [Con07] Drosophila 12 Genomes Consortium. Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, 450:203–208, 2007.
- [Fel81] Joseph Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, November 1981.
- [Fit71] W.M. Fitch. Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Zoology*, 20:406–416, 1971.
- [HKY85] M. Hasegawa, H. Kishino, and T. Yano. Dating the human-ape split by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22:160–174, 1985.
- [KGR<sup>+</sup>03] A.E. Kel, E. Gossling, I. Reuter, E. Cheremushkin, O.V. Kel-Margoulis, and E. Wingender. MATCH: a tool for searching transcription factor binding sites in DNA sequences. *Nucl. Acids Res.*, 31(13):3576–3579, 2003.
- [Kie53] J. Kiefer. Sequential Minimax Search for a Maximum. *Proceedings of the American Mathematical Society*, 4(3):502–506, 1953.
- [MCS07] Axel Mosig, Julian L. Chen, and Peter F. Stadler. Homology Search with Fragmented Nucleic Acid Sequence Patterns. In *WABI 2007 (R. Giancarlo & S. Hannehalli, eds.)*, pages 335–345, 2007.
- [MGSS07] Axel Mosig, Meng Guofeng, Brbel M. R. Stadler, and Peter F. Stadler. Evolution of the Vertebrate Y RNA Cluster. *Th Biosci.*, 126:9–14, 2007.
- [MM04] Scott McGinnis and Thomas L. Madden. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucl. Acids Res.*, 32(suppl2):W20–25, 2004.

- [MSS06] Axel Mosig, Katrin Sameith, and Peter F. Stadler. *fragrep*: Efficient Search for Fragmented Patterns in Genomic Sequences. *Geno. Prot. Bioinfo.*, 4:56–60, 2006.
- [Nie] Morten Nielsen. EasyPred web server: <http://www.cbs.dtu.dk/biotools/EasyPred/>. website.
- [PBM<sup>+</sup>02] Tal Pupko, Rachel E Bell, Itay Mayrose, Fabian Glaser, and Nir Ben-Tal. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, 18 Suppl 1:S71–S77, 2002.
- [SCH<sup>+</sup>09] Peter F. Stadler, Julian J.-L. Chen, Jörg Hackermüller, Steve Hoffmann, Friedemann Horn, Phillip Khaitovich, Antje K. Kretzschmar, Axel Mosig, Sonja J. Prohaska, Xiaodong Qi, Katharina Schutt, and Kerstin Ullmann. Evolution of Vault RNAs. *Mol. Biol. Evol.*, 2009. accepted.
- [Sto00] Gary D. Stormo. DNA binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, 2000.
- [Tav86] S Tavaré. Some probabilistic and statistical problems on the analysis of DNA sequences. *Lect. Math. Life Sci.*, 17:57–86, 1986.
- [TBF<sup>+</sup>07] The Athanasius F. Bompfinewerer RNA Consortium., Rolf Backofen, Christoph Flamm, Claudia Fried, Guido Fritsch, Jörg Hackermüller, Jana Hertel, Ivo L. Hofacker, Kristin Missal, Sonja J. Mosig, Axel Prohaska, Domininc Rose, Peter F. Stadler, Andrea Tanzer, Stefan Washietl, and Will Sebastian. RNAs Everywhere: Genome-Wide Annotation of Structured RNAs. *J. Exp. Zool. B: Mol. Dev. Evol.*, 308B:1–25, 2007.
- [Tho04] Joseph W. Thornton. Resurrecting ancient genes: experimental analysis of extinct molecules. *Nat Rev Genet*, 5(5):366–375, May 2004.
- [VS93] Martin Vingron and Peter R. Sibbald. Weighting in sequence space: A comparison of methods in terms of generalized sequences. *Proc. Natl. Acad. Sci. USA*, 90:8777–8781, 1993.
- [XMQ<sup>+</sup>08] Mingyi Xie, Axel Mosig, Xiaodong Qi, Yang Li, Peter F. Stadler, and Julian J.-L. Chen. Size Variation and Structural Conservation of Vertebrate Telomerase RNA. *J. Biol. Chem.*, 283:2049–2059, 2008.
- [Yan07] Ziheng Yang. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol*, 24(8):1586–1591, 2007.
- [ZJ97] Nei M. Zhang J. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *J Mol Evol*, 44:S139–46, 1997.