Table 3: DIALIGN alignment scores for anchored and non-anchored alignment of five reference test sets from BAliBASE. As anchor points, we used the so-called *core-blocks* in BAliBASE, thereby enforcing biologically correct alignments of the input sequences. The figures in the first and second line refer to the sum of DIALIGN alignment scores of all protein families in the respective reference set. Line four contains the number of sequence sets where the anchoring *improved* the alignment score together with the total number of sequence sets in this reference set. Our test runs show that on these test data, biologically meaningful alignments do not have higher DIALIGN scores than alignments produced by the default version of our program.

	Alignment scores					
	Ref1	Ref2	Ref3	Ref4	Ref5	Total
non-anchored	53,613	269,009	283,273	36,515	29,214	671,624
anchored	53,417	265,966	283,136	36,611	29,257	668,387
ratio	0.996	0.988	0.999	1.002	1.001	0.995
score improved	23/82	13/23	4/23	6/16	4/12	50/156

alignment was 91% while the column score was 90% as 18 out of 20 columns of the core blocks were correctly aligned. As was generally the case for BAliBASE, the *DIALIGN score* of the (biologically meaningful) anchored alignment was lower than the score of the (biologically wrong) default alignment. The DIALIGN score of the anchored alignment was 9.82 compared with 11.99 for the non-anchored alignment, so here the score of the anchored alignment was around 18 percent below the score of the non-anchored alignment.

Anchored alignments for phylogenetic footprinting

Evolutionarily conserved regions in non-coding sequences represent a potentially rich source for the discovery of gene regulatory regions. While functional elements are subject to stabilizing selection, the adjacent non-functional DNA evolves much faster. Therefore, blocks of conservation, so-called phylogenetic footprints, can be detected in orthologous non-coding sequences with low overall similarity by comparative genomics [39]. Alignment algorithms, including DIALIGN, were advocated for this task. As the example in the previous section shows, however, anchoring the alignments becomes a necessity in applications to large genomic regions and clusters of paralogous genes. While interspersed repeats are normally removed ("masked") using e.g. *RepeatMasker*, they need to be taken into account in the context of phylogenetic footprinting: if a sequence motif is conserved hundreds of millions of years it may well have become a regulatory region even if it is (similar to) a repetitive sequence in some of the organisms under consideration [40].

The phylogenetic footprinting program *TRACKER* [41] was designed specifically to search for conserved non-coding sequences in large gene clusters. It is based on a similar philosophy as segment based alignment algorithms. The TRACKER program computes pairwise local alignments of all input sequences using BLASTZ [42] with non-stringent settings. BLASTZ permits alignment of long genomic sequences with large proportions of neutrally evolving regions. A post-processing step aims to remove simple repeats recognized at their low sequence complexity and regions of low conservation. The resulting list of pairwise alignments is then assembled into clusters of partially overlapping regions. Here the approach suffers from the same problem as DIALIGN, which is, however, resolved in a different way: instead of producing a single locally optimal alignment, TRACKER lists all maximal compatible sets of pairwise alignments. For the case of Figure 1(C), for instance, we obtain both $M_1^{(1)} M_2 M_3$ and $M_1^{(2)} M_2 M_3$. Since this step is performed based on the overlap of sequence intervals without explicitly considering the sequence information at all, TRACKER is very fast as long as the number of conflicting pairwise alignments remains small. In the final step DIALIGN is used to explicitly calculate the multiple sequence alignments from the subsequences that belong to individual clusters.

For the initial pairwise local alignment step the search space is restricted to orthologous intergenic regions, parallel strands and chaining hits. Effectively, TRACKER thus computes alignments anchored at the genes from BLASTZ fragments.

We have noticed [43] that DIALIGN is more sensitive than TRACKER in general. This is due to detection of smaller and less significant fragments with DIALIGN compared to the larger, contiguous fragments returned by BLASTZ. The combination of BLASTZ and an anchored version of DIALIGN appears to be a very promising approach for phylogenetic footprinting. It makes use of the alignment specificity of BLASTZ and the sensitivity of DIALIGN. A combination of anchoring at appropriate genes (with maximal weight) and BLASTZ hits (with smaller weights proportional e.g. to – log *E* values) reduces the CPU requirements for the DIALIGN alignment by more than an order of magnitude. While this is still much slower