ters is quite obvious, there is usually no *direct* way of influencing the outcome of an alignment program.

Automated alignment methods are clearly necessary and useful where large amounts of data are to be processed or in situations where no additional expert information is available. However, if a researcher is familiar with a specific sequence family under study, he or she may already know certain parts of the sequences that are functionally, structurally or phylogenetically related and should therefore be aligned to each other. In situations where automated programs *fail* to align these regions correctly, it is desirable to have an alignment method that would accept such user-defined homology information and would then align the remainder of the sequences automatically, respecting these user-specified *constraints*.

The interactive program *MACAW* [11] can be used for semi-automatic alignment with user-defined constraints; similarly the program *OWEN* [12,13] accepts anchor points for pairwise alignment. Multiple-alignment methods accepting pre-defined constraints have also been proposed by Myers *et al.* [14] and Sammeth *et al.* [15]. The multi-alignment program DIALIGN [16,17] has an option that can be used to calculate alignments under user-specified constraints. Originally, this program feature has been introduced to reduce the alignment search space and program running time for large genomic sequences [18,19]; see also [20]. At *Göttingen Bioinformatics Compute Server (GOBICS)*, we provide a user-friendly web interface where anchor points can be used to guide the multiple alignment procedure [21]. Herein, we describe our anchored-alignment approach in detail using a previously introduced set-theoretical alignment concept. We apply our method to genomic sequences of the *Hox* gene clusters. For these sequences, the default version of DIALIGN produces serious mis-alignments where entire genes are incorrectly aligned, but meaningful alignments can be obtained if the known gene boundaries are used as anchor points.

In addition, our anchoring procedure can be used to obtain information for the further development of alignment algorithms. To improve the performance of automatic alignment methods, it is important to know what exactly goes wrong in those situations where these methods fail to produce biologically reasonable alignments. In principle, there are two possible reasons for failures of alignment programs. It is possible that the underlying *objective function* is 'wrong' by assigning high numerical scores to biologically meaningless alignments. But it is also possible that the objective function is 'correct' – i.e. biologically correct alignments have numerically optimal scores -and the employed heuristic *optimisation algorithm* fails to return mathematically optimal or near-optimal alignments. The anchoring approach that we implemented can help to find out which component of our alignment program is to blame if automatically produced alignments are biologically incorrect.

One result of our study is that anchor points can not only improve the *biological* quality of the output alignments but can in certain situations lead to alignments with significantly higher *numerical* scores. This demonstrates that the heuristic optimisation procedure used in DIALIGN may produce output alignments with scores far below the optimum for the respective data set. The latter result has important consequences for the further development of our alignment approach: it seems worthwile to develop more efficient algorithms for the optimisation problem that arises in the context of the DIALIGN algorithm. In other situations, the numerical scores of biologically correct alignments turned out to be below the scores of biololgically wrong alignments returned by the non-anchored version of our program. Here, improved optimisation functions will not lead to biologically more meaningful alignments. It is therefore also promising to develop improved objective function for our alignment approach.

Alignment of tandem duplications

There are many situations where automated alignment procedures can produce biologically incorrect aligments. An obvious challenge are *distantly* related input sequences where homologies at the primary sequence level may be obscured by spurious random similarities. Another notorious challenge for alignment programs are *duplications* within the input sequences. Here, *tandem duplications* are particularly hard to align, see e.g. [22]. Specialised software tools have been developed to cope with the problems caused by sequence duplications [23]. For the segment-based alignment program DIALIGN, the situation is as follows. As described in previous publications, the program constructs pairwise and multiple alignments from pairwise local sequence similarities, so-called *fragment alignments or fragments* [17,16]. A fragment is defined as an un-gapped pair of equal-length segments from two of the input sequences. Based on statistical considerations, the program assigns a *weight score* to each possible fragment and tries to find a consistent collection of fragments with maximum total score. For pairwise alignment, a *chain* of fragments with maximum score can be identified [24]. For multiple sequence sets, all possible pairwise alignments are performed and fragments contained in these pairwise alignments are integrated *greedily* into a resulting multiple alignment.

As indicated in Figure 1, tandem duplications can create various problems for the above outlined alignment approach. In the following, we discuss two simple exam-