Conclusion

Automated alignment procedures are based on simple algorithmical rules. For a given set of input sequences, they try to find an alignment with maximum score in the sense of some underlying objective function. The two basic questions in sequence alignment are therefore (*a*) to define an meaningful objective function and (*b*) to design an efficient optimisation algorithm that finds optimal or at least near-optimal alignments with respect to the chosen objective function. Most multi-alignment programs are using *heuristic* optimisation algorithms, i.e. they are, in general, not able to find the mathematically optimal alignment with respect to the objective function. An objective function for sequence alignment should assign *numerically* high scores to *biologically* meaningful alignments. However, it is clearly not possible to find a *universally* applicable objective function that would give highest numerical scores to the biologically correct alignments in all possible situations. This is the main reason why alignment programs may fail to produce biologically reasonable output alignments. In fact, the impossibility to define a universal objective function constitutes a fundamental limitation for *all* automated alignment algorithms.

Often a user is already familiar with a sequence family that he or she wants to align, so some knowledge about existing sequence homologies may be available. Such expert knowledge can be used to direct an otherwise automated alignment procedure. To facilitate the use of expert knowledge for sequence alignment, we proposed an *anchored alignment* approach where known homologies can be used to restrict the alignment search space. This can clearly improve the quality of the produced output alignments in situations where automatic procedures are not able to produce meaningful alignments. In addition, alignment anchors can be used to reduce the program running time. For the *Hox* gene clusters that we analyzed, the non-anchored version of DIALIGN produced serious mis-alignments. We used the known gene boundaries as anchor points to guarantee a correct alignment of these genes to each other.

There are two possible reasons why automated alignment procedures may fail to produce biologically correct alignments, (*a*) The chosen objective function may not be in accordance with biology, i.e., it may assign mathematically high scores to biologically wrong alignments. In this case, even efficient optimisation algorithms would lead to meaningless alignments. (*b*) The mathematically optimal alignment is biologically meaningful, but the employed heuristic optimisation procedure is not able to find the alignment with highest score. For the further development of alignment algorithms, it is crucial to find out which one of these reasons is to blame for mis-alignments produced by existing software programs. If (*a*) is often observed for an alignment program, efforts should be made to improve its underlying objective function. If (*b*) is the case, the biological quality of the output alignments can be improved by using a more efficient optimisation algorithm. For DIALIGN, it is unknown how close the produced alignments come to the numerically optimal alignment – in fact, it is possible to construct example sequences where DIALIGN's greedy heuristic produces alignments with arbitrarily low scores compared with the possible optimal alignment.

In the Fugu example, Figure 2 and 3, the *numerical* alignment score of the (anchored) correct alignment was 13% below the score of the non-anchored alignment. All sequences in Figure 2 and 3 contain only subsets of the 13 *Hox* paralogy groups, and different sequences contain different genes. For such an extreme data set, it is unlikely that any reasonable objective function would assign an optimal score to the biologically correct alignment. Here, the problem is that sequence similarity no longer coincides with biological homology. The only way of producing good alignments in such situations is *to force* a program to align certain known homologies to each other. With our anchoring approach we can do this, for example by using known gene boundaries as *anchor points*.

For the BAliBASE benchmark data base, the total score of the (biologically meaningful) anchored alignments was also below the score of the (biologically wrong) non-anchored default alignments.

This implies, that improved optimisation algorithms will not lead to biologically improved alignments for these sequences. In this case, however, there is some correspondence between sequence similarity and homology, so one should hope that the performance of DIALIGN on these data can be improved by to designing better objective functions. An interesting example from BAliBASE is shown in Figure 4. Here, the non-anchored default version of our program produced a complete mis-alignment. However, it was sufficient to enforce the correct alignment of one *single* column using corresponding anchor points to obtain a meaningful alignment of the entire sequences where not only the one anchored column but most of the three core blocks are correctly aligned. This indicates that the correct alignment of the core blocks corresponds to a *local maximum* in the alignment landscape.

In contrast, in the teleost *HoxA* cluster example the numerical score of the anchored alignment was around 15% *above* the score of the non-anchored alignment. This demonstrates that the greedy optimisation algorithm used by DIALIGN can lead to results with scores far below the optimal alignment. In such situations, improved optimisation algorithms may lead not only to mathematically