Poster H-50

M-Coffee: combining multiple sequence alignment methods with T-Coffee.

Authors:
Iain Wallace (*University College Dublin,Ireland*)
Orla OSuillivan (*University College Dublin,Ireland*)
Desmond Higgins (*University College Dublin,Ireland*)
Cedric Notredame (*Laboratoire Information Génomique et Structurale, CNRS Marseille,France*)

Short Abstract: We introduce M-Coffee, an extension of T-Coffee, a meta-method for assembling multiple sequence alignments (MSA) by combining the output of several individual methods into one single MSA. M-Coffee outperforms all the individual methods on three major reference datasets: HOMSTRAD, Prefab and Balibase.

Long Abstract:

Introduction

The multiple alignment of DNA or protein sequences is one of the most commonly used techniques in sequence analysis. Multiple alignments constitute a necessary pre-requisite in phylogeny, remote homologue detection and structure prediction. Until recently the choice for building multiple sequence alignments (MSAs) was limited to a handful of packages but a recent increase in genomic data has fuelled the development of many novel methods arguably more accurate and faster than the older ones. In practice this widened choice has also made it harder to objectively choose the appropriate method for a specific problem. More than 50 MSA methods have been described over the last 10 years (Medline, January 08, 2006), with no less than 20 new publications in 2005 alone. The complexity and variety of these algorithms and the fact than none provides a definite answer to the problem makes it almost impossible to tell them apart from a theoretical point of view.

The description of a complex problem partially solved by several more or less different methods calls for comparisons with other similar situations in computational biology like secondary structure and gene predictions. In these contexts, Meta-methods, or Jury-based methods have often proven to be superior to the constitutive methods. However, in the case of gene or structure predictions, the output is relatively easy to combine into the intersection or union of individual predictions. Such a combination protocol is harder to define when it comes to MSAs where each pair of aligned residues constitutes an element of prediction. Fortunately, consistency-based objective functions provide an elegant and simple solution to the problem of averaging several alignments into one meaningful consensus. Given a collection of alternative alignments, consistency-based objective functions define the optimal alignment as the one having the highest level of consistency with the collection. It is realistic to consider this optimally consistent alignment as some sort of consensus. This approach, first described by Bucka-Lassen et al. for the combination of alternative DNA alignments, is the core of the T-Coffee algorithm. While any consistency-based packages currently available would probably be equally well suited to the combination of MSAs, T-Coffee bears

the advantage of having been specifically designed for that purpose thanks to the concept of a library. T-Coffee does not explicitly align sequences but compiles libraries based on externally produced alignments. During the alignment process, the libraries are combined into the final MSA. Originally generated using ClustalW and Lalign, the libraries can also be produced by structural alignment packages or any sequence alignment program, pairwise or multiple. In this work, we took this concept much further and showed that T-Coffee can easily combine up to 15 alternative MSAs of the same sequences. We call this meta-mode M-Coffee and using several well-known benchmarks, we show that M-Coffee is the most accurate and flexible MSA meta-method described so far.

Results

Our first task was to determine how the 15 MSA methods considered here should be combined into one consensus alignment. Given 15 methods, one should consider either defining an optimal subset or devising a weighting scheme that makes it possible to combine all the methods at once. Our first attempt was to use a greedy procedure in order to define an optimal subset of methods. Methods were ranked according to their overall accuracy on the 233 HOMSTRAD reference datasets and the order thus defined was used to define subsets of methods used within M-Coffee. Results are shown on Figure 1, where subset 1 only contains the best method (ProbCons), subset 2 contains the best and the second best (ProbCons + Muscle 6), and so on. The graph clearly shows a peak, which is significantly better than the point before it (Wilcoxon $P < 0.001$), suggesting that an accumulation of low accuracy methods eventually affects the overall results. On the other hand, the graph also indicates that except for the two first subsets, the accuracy of M-Coffee is clearly higher than any of the constituting methods, thus establishing the efficiency of the combination.

The degradation in accuracy when very similar methods are added, like the MAFFT family of programs (FFTNSI, FFTNS2, FFTNS1 and so on), is not surprising when considering the underlying principle of consistency. Consistency is only useful as an accuracy indicator when methods are unlikely to commit exactly the same error. However, this assumption is no longer true when nearly identical methods are being combined. When this happens, incorrect alignment portions find their way into the final
model simply because they appear highly consistent to the T-Coffee algorithm.

We applied various weighting schemes but found that they did not appear to properly address the problem of method redundancy, and the overall results suggest a need for some crude and discrete filtering. We eventually considered that arbitrary code setting (e.g. choosing between alignments with equal scores) could be one of the reasons for misleading consistency between packages of the same groups. This lead us to hand pick one method per developer (the most accurate) and use the resulting subsets to run our tests. The eight selected methods were POA-global, Dialign-T, ClustalW, PCMA, FINSI, T-Coffee, Muscle v6 and ProbCons. This combination of methods will be called M-Coffee8. Results are shown on Figure 2. Interestingly, M-Coffee8 outperforms any of the constitutive method all along the combination process, thus suggesting an always beneficial combination. Figure 2 also shows that M-Coffee8 is more accurate than ProbCons, even before inclusion of that method.

Figure 1: CS after combining multiple alignment methods with T-Coffee.

Figure 2: M-Coffee8. The top line (closed diamonds) is the CS on the HOMSTRAD benchmark after combining multiple alignments using only one method per developer. The bottom line (closed squares) is the default performance for each method on the benchmark.