HMM-Kalign: a tool for generating sub-optimal HMM alignments

Emmanuelle Becker*, Aurélie Cotillard, Vincent Meyer, Hocine Madaoui and Raphaël Guérois*

CEA, iBiTecS, URA 2096, SB²SM, Laboratoire de Biologie Structurale et Radiobiologie, Gif sur Yvette, F-91191 France

Received on April 10, 2007; revised on September 10, 2007; accepted on September 26, 2007

Advance Access publication October 6, 2007

Associate Editor: Anna Tramontano

ABSTRACT

Summary: Recent development of strategies using multiple sequence alignments (MSA) or profiles to detect remote homologies between proteins has led to a significant increase in the number of proteins whose structures can be generated by comparative modeling methods. However, prediction of the optimal alignment between these highly divergent homologous proteins remains a difficult issue. We present a tool based on a generalized Viterbi algorithm that generates optimal and sub-optimal alignments between a sequence and a Hidden Markov Model. The tool is implemented as a new function within the HMMER package called hmmpalign.


Contact: raphael.guerois@cea.fr; emmanuelle.becker@gmail.com

Supplementary information: Supplementary data are available at Bioinformatics online.

The present work aims at automatically exploring the alignment space in the neighborhood of the optimal sequence alignment (OSA) in order to find an alignment closer to the structural alignment than the OSA.

The sequence alignment space in the neighborhood of the OSA has been quite extensively explored in the context of pairwise sequence alignments. Waterman, (1983) proposed an algorithm derived from the standard Sellers algorithm to determine all the pairwise alignments whose scores are within a range \( e \) of the OSA’s score. Later and still dealing with pairwise sequence alignments, Saqi and Sternberg (Saqi and Sternberg, 1991) proposed a heuristic known as the Iterative Elimination Method, based on the progressive perturbation of the distance matrix. Another method to generate alternative pairwise sequence alignments has been introduced by Zuker, (1991).

With the rise of sequence-profile, sequence-HMM and more recently profile-profile and HMM–HMM alignments, these algorithmic developments were less studied. However, although progress has been made especially for the detection of remote homology, the alignment of sequences sharing less than 25% of sequence identity is still problematic in the context of comparative modeling. Based on this observation, some articles (Chivian and Baker, 2006; Jaroszewski et al., 2002; John and Sali, 2003) re-introduced the idea of generating alternative alignments by using heuristics such as a parametric approach (Chivian and Baker, 2006) coupled with Saqi and Sternberg’s Iterative Elimination Method (Jaroszewski et al., 2002), or a genetic algorithm (John and Sali, 2003).

In this work, we explore the possibility of generating alternative alignments in the context of alignments obtained using Hidden Markov Models, such as HMMER (Eddy, 1996) or SAM (Karpplus et al., 2005). Instead of heuristics, HMM-Kalign generates the exact neighborhood of the OSA.

The Viterbi algorithm is classically used to align a sequence \( s_{\text{obs}} \) to a profile HMM and consists in finding the sequence of states that maximizes the emission probability of \( s_{\text{obs}} \) (Viterbi, 1967). To generate alternative alignments in the neighborhood of the OSA, one solution is to use a generalized Viterbi algorithm that precisely determines the \( k \)-best sequences of states that maximize the emission of \( s_{\text{obs}} \). This generalization of the Viterbi algorithm has been used in the field of speech recognition and elegant variants have been developed recently that fasten the process (Huang and Chiang, 2005). We implemented and included the generalized Viterbi algorithm in the program HMMER (Eddy, 1996).

1 GENERATING SUB-OPTIMAL ALIGNMENTS

To use the hmmpalign command, two files are required:

- \(<\text{MSA}>\), that contains a multiple sequence alignment (derived for example from the alignment of structural templates);
- \(<\text{sequences}>\), that contains two sequences in fasta format: (i) the sequence to be aligned, (ii) one sequence from the \(<\text{MSA}>\) file that may be used as a template to further build a model of the first sequence.

To build the HMM, it is possible to use the classical command:

$ ./hmm_build <hmm file> <MSA> (command 1)

although our results show that within highly divergent families, it is more effective to drive explicitly the HMM architecture with respect to the conservation of the secondary structures (details in Supplementary Material 1). This is possible via the command:

$ ./hmm_build --hand <hmm file> <MSA> (command 2)

where the \(<\text{MSA}>\) file contains an additional line with symbols ‘-‘ and ‘x‘ encoding for the positions of insertions

*To whom correspondence should be addressed.
and match states, respectively. After building the HMM, the command to generate $k$ alignments is:

\[
\text{hmmkalign } k \text{ hmm file }
\]

The OSA classically generated with HMMER corresponds to the alignment with the best score ($K = 1$) (cf. command 1).

Exploration can be targeted to specific regions. For a sequence $s_{\text{obs}} = s_1 \ldots s_T$ in which only the region $s_i \ldots s_j$ is to be sampled, add a hybrid sequence in the $\text{MSA}$ file, that contains the ‘anchors’ $s_1 \ldots s_{i-1}$ and $s_{j+1} \ldots s_T$ and insertions ‘-‘ symbols instead of $s_i \ldots s_j$.

2 TESTING PROCEDURE

We studied 115 alignments from 22 highly divergent protein families, sharing on average $525\%$ identity (Supplementary Material 2). These alignments were extracted from the HOMSTRAD database (Stebbings and Mizuguchi, 2004). The following procedure was applied: (1) exclude the test sequence from the multiple structural alignment; (2) build two distinct HMMs with commands 1 and 2; (3) get 20 sub-optimal alignments of the excluded sequence on both HMMs and (4) evaluate with respect to the structural alignment.

3 RESULTS FOR THE 115 TEST CASES

Generating only 40 sub-optimal alignments we found that in 95 of the 115 test cases, at least one sub-optimal alignment had a $Q_{\text{mod}}$ better than the OSA. For 26 of them, the $Q_{\text{mod}}$ increased by more than 0.10 (Supplementary Material 4). The alternative alignments generated by $\text{hmmkalign}$ were also found to be of greater interest than the ones generated with heuristic approaches (Supplementary Material 4). These results highlight that targeted sampling of the sequence alignment space in the neighborhood of the OSA by $\text{hmmkalign}$ is efficient in generating optimized alignments and thereby better models.

4 EXAMPLE WITHIN THE THIOREDOXIN FAMILY

The thioredoxin family contains small enzymes that are involved in redox reactions. Their sequences are on average 100 amino acids long and highly divergent (17\% sequence identity on average), while their three-layers sandwich fold is conserved. Aligning the sequence of the oxidized bacteriophage T4 glutaredoxin with the other members of the family is a difficult task. As a matter of fact, the OSA (Fig. 1b) is far from the structural alignment (ratio of correctly aligned positions $Q_{\text{mod}} = 0.50$).

First, we studied the 20 sub-optimal alignments produced when the HMM is built with command 1 (Fig. 1b). The alignment can be divided in two parts: the first 63 amino acids, whose positions are extremely variable, and the last 24 amino acids that are not shifted. Not surprisingly, the least varying positions along the sampled alignments correlate with the correctly aligned ones. Within the sub-optimal alignments, alignments $K = 12$, $K = 15$ and $K = 16$, are substantially better than the OSA ($Q_{\text{mod}} = 0.79$).

We then studied the 20 sub-optimal alignments produced when HMM architecture is explicitly driven by secondary structure conservation (cf. command 2). Alignments with $Q_{\text{mod}}$ reaching 0.89 were obtained (three of them are shown in Fig. 1c).
Homology models of the oxidized bacteriophage T4 glutaredoxin were constructed with the OSA and all the sub-optimal alignments. As illustrated in Figure 1d, the root mean square deviation between the native structure and the models are much smaller with models produced with the sub-optimal alignments ($K = 16$ or $K = 7$) than with models produced with the OSA.

ACKNOWLEDGEMENT
This work is partly funded by the ACI IMPBIO SpIDER.

Conflict of Interest: none declared.

REFERENCES