Bioinformatics, 31(23), 2015, 3850–3852
doi: 10.1093/bioinformatics/btv441
Advance Access Publication Date: 30 July 2015
Applications Note

OXFORD

Structural bioinformatics

HHalign-Kbest: exploring sub-optimal alignments for remote homology comparative modeling

Jinchao Yu[1], Geraldine Picord[2,3,4], Pierre Tuffery[2,3,4] and Raphael Guerois[1,*]

[1]Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, University Paris-Saclay, CEA-Saclay, 91191 Gif-sur-Yvette, [2]INSERM U973, MTi, F-75205 Paris, [3]Université Paris Diderot, Sorbonne Paris Cité F-75205 Paris and [4]Ressource Parisienne en Bioinformatique Structurale, F-75205 Paris, France

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on July 10, 2014; revised on June 22, 2015; accepted on July 21, 2015

Abstract

Motivation: The HHsearch algorithm, implementing a hidden Markov model (HMM)-HMM alignment method, has shown excellent alignment performance in the so-called twilight zone (target-template sequence identity with ~20%). However, an optimal alignment by HHsearch may contain small to large errors, leading to poor structure prediction if these errors are located in important structural elements.

Results: HHalign-Kbest server runs a full pipeline, from the generation of suboptimal HMM-HMM alignments to the evaluation of the best structural models. In the HHsearch framework, it implements a novel algorithm capable of generating *k*-best HMM-HMM suboptimal alignments rather than only the optimal one. For large proteins, a directed acyclic graph-based implementation reduces drastically the memory usage. Improved alignments were systematically generated among the top *k* suboptimal alignments. To recognize them, corresponding structural models were systematically generated and evaluated with Qmean score. The method was benchmarked over 420 targets from the SCOP30 database. In the range of HHsearch probability of 20–99%, average quality of the models (TM-score) raised by 4.1–16.3% and 8.0–21.0% considering the top 1 and top 10 best models, respectively.

Availability and implementation: http://bioserv.rpbs.univ-paris-diderot.fr/services/HHalign-Kbest/ (source code and server).

Contact: guerois@cea.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Automated protein structure prediction has been widely used for biological research in recent years. HHsearch is among the fastest and most accurate tool for remote homology detection and target-template alignment by using a hidden Markov model (HMM) pairwise alignment method (Hildebrand *et al.*, 2009; Söding, 2005). Despite its high performance, an optimal alignment often contains small to large errors, especially in the position and length of gaps. Usually, structural biologists correct these errors manually using back and forth examination of the alignments and of the resulting structural models. This task becomes particularly tedious when target-template sequence identities are low and hampers the development of fully automated pipelines able to generate accurate structural models.

In this work, we present a tool integrating suboptimal technique within the Viterbi algorithm of HHsearch program to generate suboptimal alignments. This interest of suboptimal exploration has previously been shown for the prediction of membrane protein topology (Brown and Golod, 2010) and in sequence-HMM pairwise alignments (Becker *et al.*, 2007). We first studied 26 targets from the CASP10 TBM dataset (template-based modeling) (Huang *et al.*, 2014) to test the algorithms. The SCOP30 database (sequence identity in the range of 5–30%, Fox *et al.*, 2014) was then used as a benchmark to define the conditions in which the strongest increase in model accuracy could be expected from the use of the HHalign-Kbest server.

2 Methods

Our *k*-best Viterbi algorithms were implemented in the latest package HHsuite 2.0.16, in which four files were modified: hhhit.C implementing the algorithms and hhhit.h, hhalign.C and hhdecl.h in which the parameters and configurations were adjusted.

To set up the method, two sets of target domains were chosen out of the 99 real targets in the CASP10 TBM category. Set A contains target domains containing segment(s) that HHpredA did not predict well but some other tools did (13 targets). As a control, set B contains 13 target domains randomly selected from the set for which HHpredA performed very well (65 targets) (Supplementary Method S1). The SCOP30 benchmark set was generated using the SCOP30 database (Fox *et al.*, 2014) (sequence identity in the range 5–30%), we randomly selected 70 non-redundant superfamily pairs in six ranges of HHsearch probabilities ((0,20%); [20%,90%); [90%,95%); [95%,99%); [99%,100%); [100%]) (Supplementary Method S2, the results for the range (0, 20%) only discussed in the Supplementary Materials).

3 Results

The Viterbi algorithm (for pair HMMs) of HHsearch is a hybrid between standard Viterbi and pairwise alignment algorithms and uses a dynamical programming matrix of three dimensions: target HMM length (noted as m), template HMM length (n) and the number of pair-state types ($N = 5$) (Söding, 2005). To obtain k-best paths, a natural extension is to add another dimension of length k to this three-dimensional matrix. This fourth dimension stores an ordered top-k scores and can be calculated by an N-way merge sort efficiently. This matrix-based k-best Viterbi algorithm has both a theoretical space complexity and time complexity of $O(N*k*m*n)$.

Since the memory usage limits the application of the matrix-based method, we developed the directed acyclic graph-based algorithm. Based on the same principle, dynamical allocation of memory is used, which gives an opportunity to prune useless paths and nodes. This method requires dramatically less memory while taking only slightly longer time (Supplementary Results S1 and S2).

CASP10 sets were used to evaluate whether suboptimal alignments could correct misaligned segment(s) in set A targets without degrading the accuracy of nearly perfect alignments of set B. We compared (sub)optimal alignments to a reference structural alignment calculated from the superimposition between the template and query PDB structures using TM-align (Zhang and Skolnick, 2005). Qloc score was used to report the accuracy of an alignment [from 0 to 1 indicating from wrong to correct, see Supplementary Method S4 (Sadreyev and Grishin, 2003)]. Figure 1A shows an example of suboptimal alignments for a target CASP10 set A (see Section 2), where the suboptimal N.398 (Qloc = 0.860) corrected the second beta sheet (red), while N.485 (Qloc = 0.835) corrected the first one (blue) perfectly in comparison with the Viterbi optimal alignment (Qloc = 0.722) (see details in Supplementary Fig. S6). Using a procedure combining Modeller (Eswar *et al.*, 2006) to generate models for each suboptimal alignment and Qmean Zscore to evaluate model quality (Benkert *et al.*, 2011), the suboptimal alignments in Figure 1A could be discriminated (Supplementary Fig. S6C). N.398 and N.485 led to better models with TM-score (average TM-score of 20 models for each (sub)optimal alignment) 0.646 and 0.652, respectively, in comparison to Viterbi alignment TM-score 0.585 (Supplementary Fig. S7B). No loss of accuracy was observed for set B (Supplementary Fig. S5).

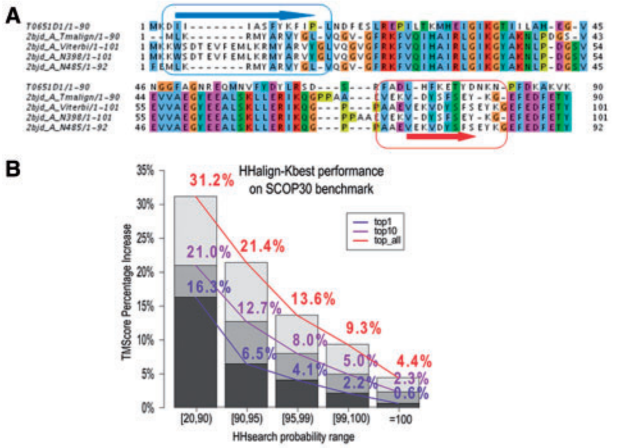

Fig. 1. (**A**) Example of suboptimal alignments obtained for CASP10 target T0651D1 over the template 2bjd_A. The positions of two strands were correctly predicted in suboptimal alignments N.398 and N.485 but not in the optimal Viterbi alignment. TM-align provides a gold standard to which these alignments can be compared. (**B**) Average percentage increase of TM-score for a domain modelled on template of the same superfamily as referenced by SCOP30. Five categories were defined according to the HHsearch probability containing 70 non-redundant cases each. Results are reported for the first model (top 1 in purple), the best among the first 10 models (top 10 in magenta) and the best among all suboptimal models ('top_all' in red)

To benchmark HHalign-Kbest, 70 pairs of models/templates extracted from SCOP30 database were randomly selected in every six ranges of HHsearch probabilities (see Section 2). Pairs could be *a posteriori* divided into hard and easy cases depending on the TM-score of the Viterbi model, below or above 50%, respectively (Zhang and Skolnick, 2004). TM-scores were improved by 4.1–6.5% (top 1) and 8.0–12.7% (top 10) for HHsearch probabilities in the range 90–99%. The improvements increased up to 16.3% (top 1) and 21.0% (top 10) for the range 20–90%. Above 99%, there were almost no hard cases, and HHalign-Kbest could still make minor improvements although it is not in the scope of use of the method (Fig. 1B, Supplementary Fig. 12S, Table 3). More details about scores, results analysis and server usage are provided in Supplementary Figures S5–S11 and Supplementary Discussion. As a general algorithm, HHalign-Kbest may be integrated in alternative frameworks for model generation such as M4T to consider multiple templates (Fernandez-Fuentes *et al.*, 2007) and with alternative evaluation scores to better recognize the best models from the 'top_all' k-best suboptimal alignments ($k = 500$).

Funding

This work was supported by ANR-IAB-2011 'BIP-BIP' and IDEX Paris Saclay—IDI 2013.

Conflict of Interest: none declared.

References

Becker,E. *et al*. (2007) HMM-Kalign: a tool for generating sub-optimal HMM alignments. *Bioinformatics*, **23**, 3095–3097.

Benkert,P. *et al*. (2011) Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics*, **27**, 343–350.

Brown,D.G. and Golod,D. (2010) Decoding HMMs using the k best paths: algorithms and applications. *BMC Bioinformatics*, **11**(Suppl 1), S28.

Eswar,N. *et al*. (2006) Comparative protein structure modeling using Modeller. *Curr. Protoc. Bioinform.*, 5–6.

Fernandez-Fuentes,N. *et al*. (2007) M4T: a comparative protein structure modeling server. *Nucleic Acids Res.*, **35**(Web Server issue), W363–W368.

Fox,N.K. *et al*. (2014) SCOPe: structural classification of proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.*, **42**(Database issue), D304–D309.

Hildebrand,A. *et al*. (2009) Fast and accurate automatic structure prediction with HHpred. *Proteins*, **77**, 128–132.

Huang,Y.J. *et al*. (2014) Assessment of template-based protein structure predictions in CASP10. *Proteins*, **82**(Suppl. 2), 43–56.

Sadreyev,R. and Grishin,N. (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J. Mol. Biol.*, **326**, 317–336.

Söding,J. (2005) Protein homology detection by HMM–HMM comparison. *Bioinformatics*, **21**, 951–960.

Zhang,Y. and Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702–710.

Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.