Bioinformatics, 31(11), 2015, 1836–1838
doi: 10.1093/bioinformatics/btv054
Advance Access Publication Date: 31 January 2015
Applications Note

OXFORD

Sequence analysis

Learning HMMs for nucleotide sequences from amino acid alignments

Carlos N. Fischer[1,*], Claudia M. A. Carareto[2], Renato A. C. dos Santos[3], Ricardo Cerri[4], Eduardo Costa[5], Leander Schietgat[6] and Celine Vens[7]

[1]Department of Statistics, Applied Maths, and Computer Science, UNESP - São Paulo State University, Rio Claro, SP, Brazil, [2]Department of Biology, UNESP-São Paulo State University, São José do Rio Preto, SP, Brazil, [3]Institute of Biosciences, UNESP-São Paulo State University, Rio Claro, SP, Brazil, [4]Department of Computer Science, UFSCar-Federal University of São Carlos, São Carlos, SP, Brazil, [5]Department of Computer Science, USP-University of São Paulo, São Carlos, SP, Brazil, [6]Department of Computer Science, KU Leuven, Leuven, Belgium and [7]Department of Public Health and Primary Care, KU Leuven Kulak, Kortrijk, Belgium

[*]To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on July 7, 2014; revised on January 14, 2015; accepted on January 26, 2015

Abstract

Profile hidden Markov models (profile HMMs) are known to efficiently predict whether an amino acid (AA) sequence belongs to a specific protein family. Profile HMMs can also be used to search for protein domains in genome sequences. In this case, HMMs are typically learned from AA sequences and then used to search on the six-frame translation of nucleotide (NT) sequences. However, this approach demands additional processing of the original data and search results. Here, we propose an alternative and more direct method which converts an AA alignment into an NT one, after which an NT-based HMM is trained to be applied directly on a genome.

Contact: carlos@rc.unesp.br

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

1 Introduction

The identification of protein coding sequences is an important step in genome annotation. Profile hidden Markov models (profile HMMs) (Eddy, 1998) are probabilistic models that can efficiently search for characteristic domains of protein families in databases. For this purpose, profile HMMs can be trained from alignments of amino acid (AA) sequences, called here AA-HMMs.

As nucleotide (NT) sequences degrade more rapidly than the AA sequences they encode, AA alignments are generally more accurate than their corresponding NT ones (Abascal *et al.*, 2010; Wernersso and Pedersen, 2003). Moreover, AA alignments can be more easily obtained than NT alignments: databases like Pfam (Punta *et al.*, 2012) and CDD (Marchler-Bauer *et al.*, 2013) make available AA alignments related to many protein domains. Therefore, AA-HMMs have also been used for searching for protein domains in genomes. The usual way of doing this is to translate the genome sequence into six AA sequences according to the reading frames and apply the HMMs on each one. However, this approach demands additional treatment of both the original data and the search results, such as translating the whole genome into six AA sequences. Furthermore, depending on the tool to run the HMMs, it is necessary to cut each AA sequence into small windows and create overlapping windows to prevent loss of regions of interest, leading to a lot of overhead. The length of the windows and the overlapping regions depend on the HMM to be used, which becomes worse when using several HMMs with very different numbers of states. Moreover, the relative positions of the HMMs predictions have to be mapped onto the real positions in the genome.

In this work, we propose a method that converts AA alignments into NT ones to learn HMMs from NT, called here NT-HMMs, for searching directly on NT sequences. This method is an alternative in order to avoid all the additional data and results processing previously described. We tested our method searching for different types of protein domain in several organisms. The method was developed for using with the program HMMER (Eddy, 2009).

2 Methods

In this section, we describe the main aspects of our method; details can be found in the Supplementary Material.

2.1 Converting AA alignments into NT alignments

The conversion from an AA alignment to an NT alignment requires a coding matrix. This matrix consists of 20 rows; each row represents an AA and each column has one of the possible codons related to that AA. The set of codons for each AA must be repeated in each row for a certain number of times, obtained as follows. First, the lowest common multiple (LCM) among the quantities of codons of all AA is calculated. The LCM value is then divided by the quantity of codons of a given AA, producing the number of repetitions of its set of codons in its corresponding row. For example, for the Universal Genetic Code, the AA are coded from 1, 2, 3, 4 or 6 codons; the LCM between these values is 12. The LCM is the number of columns of the matrix. Thus, if an AA is coded from four different codons, as for example Alanine, its set of codons is repeated three times in its row of the matrix. For Tryptophan its unique codon appears repeated 12 times in its row. Using this coding matrix, in the conversion process, each AA sequence of the original alignment gives rise to 'LCM' sequences of NT (in the example, LCM = 12): to produce each resulting NT sequence, each AA of a sequence is replaced every time by one of the 'LCM' codes (codons) that appear in its corresponding row of the coding matrix, which causes an AA alignment of N sequences to produce an NT alignment of 12N sequences.

The conversion method proposed here is supported by the fact that a state sequence of an HMM is a first-order Markov chain (Eddy, 1998) but the emission probabilities for each state are calculated taking into account only one specific position of the alignment at each time. Considering this, the order of the codons inside of the respective sets is irrelevant—what is important is the number of repetitions of each codon set (indeed, the number of each codon) in each row of the conversion matrix. Furthermore, when an HMM evaluates a sequence, each AA or NT is scored by an HMM state in an independent way of the rest of the alignment for that sequence (Eddy, 1998). Thus, each AA of an alignment of AA sequences can be replaced directly by its possible codons. So, for example, there is no need to combine the possible codons of all AA for that sequence.

The conversion process ensures that the final NT alignment contains sequences with the same length, which is required by HMMER. Also, this ensures the proportionality between NT for each AA in relation to the three positions of a codon. Then, the NT alignment is ready for learning an HMM.

3 Results

In this section, we verify whether HMMs built on converted NT alignments are able to retrieve genome sequences related to the tested domains. We compare the results produced by corresponding NT-HMMs and AA-HMMs with each other and also compare them with domain annotations for the genomes, if they are available. Details can be found in the Supplementary Material.

3.1 Searching for domains of Ribonuclease H

Initially, we tested our method by searching for Ribonuclease H of Retrotransposons (RNase_HI_RT_Ty1 and RNase_HI_RT_Ty3) on chromosomes of *Drosophila melanogaster* present in Flybase (St Pierre *et al.*, 2014)—the AA alignments were obtained from CDD (see Supplementary Table S1) and converted into NT ones using our conversion method. An AA-HMM and an NT-HMM for each subtype of RNase H were learned using the alignments.

We compare the search results from corresponding HMMs with each other (we did not find annotations to compare with). The results show that both types of HMMs of RNase_HI_RT_Ty1 and RNase_HI_RT_Ty3 identified, respectively, the same 58 and 323 potentially full-length sequences (the ones that putatively are functional) in all tested chromosomes. Furthermore, respectively, 20 and 83 defective sequences (harboring indels—insertions or deletions—hampering the domain integrity) were equally predicted by both types of HMM. The number of unmatched predictions was very low (see Supplementary Tables S5 and S6).

3.2 Searching for phosphatases in bacteria

In the second test, we used AA-HMMs and NT-HMMs to search for phosphatase domains in genomes of bacteria (Supplementary Table S4). We were interested in specific phosphatase domains that, according to Pfam, are found in some bacteria but not in others. The AA alignments of these domains were obtained from Pfam (Supplementary Table S2), converted into NT ones and used to train the HMMs.

For the tested domains and genomes, the related AA-HMM and NT-HMM predicted virtually the same annotated sequences—the differences between the start (respectively, end) positions of corresponding predictions are always less than 5 NT. The number of false positives was very low (Supplementary Tables S7–S11).

3.3 Searching for CBM_1 and Fungal_trans in fungi

In this test, we searched for the carbohydrate-binding module (CBM_1) and fungal specific transcription factor (Fungal_trans)—their AA alignments were obtained from Pfam (Supplementary Table S3) and converted into NT ones. Four HMMs were trained using these alignments and run on the *Aspergillus fumigatus* and *A. niger* genomes. We compared the predictions of both types of HMM with annotations for the tested fungi.

The CBM_1 domain annotations for these fungi describe 17 sequences in *A. fumigatus* and 8 sequences in *A. niger*. All these sequences were predicted correctly by both NT-HMM and AA-HMM. The number of false positives was very low (Supplementary Table S12).

For Fungal_trans, 165 sequences are annotated in *A. fumigatus*: 21 of them were not predicted by any HMM; additionally, the AA-HMM missed 9 annotations, which were predicted by the NT-HMM, and the NT-HMM missed 14 other ones that were identified by the AA-HMM. In *A. niger*, 223 sequences are annotated: 20 of them were missed by both HMMs; the AA-HMM did not predict 23 other ones, identified by the NT-HMM; the NT-HMM missed 28 annotations that were predicted by the AA-HMM (Supplementary Tables S13 and S14). A possible explanation for these losses would be the presence of introns inside the domain sequences. These numbers suggest that, in situations like this, the total of correct predictions could be improved by combining the results of both types of HMM.

4 Conclusions

In this paper, we describe a method to convert an AA alignment into an NT one. Experiments on several genomes show that the NT-HMMs trained using converted alignments presented the same performance as the corresponding AA-HMMs for RNase H, phosphatases and CBM_1 domains. For Fungal_trans, the AA-HMM was slightly better than the corresponding NT-HMM. The results

also show that both types of HMM can lose domain sequences when indels or introns are present inside them. In these cases, both types of HMM could be used together to increase the number of correct predictions.

Funding

This work was supported by the Explorative Scientific Cooperation Program between São Paulo State University-UNESP/Brazil and KU Leuven/Belgium, by the São Paulo Research Foundation-FAPESP/Brazil (grant 2012/24774-2 to C.N.F. and grant 2010/10731-4 to C.M.A.C.) and by the National Council for Scientific and Technological Development-CNPq/Brazil (CNPq fellowship 306493/2013-6 to C.M.A.C.).

Conflict of Interest: none declared.

References

Abascal,F. *et al*. (2010) TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.*, **38**, W7–W13.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinf. Rev.*, **14**, 755–763.

Eddy,S.R. (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inf.*, **23**, 205–211.

Marchler-Bauer,A. *et al*. (2013) CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Res.*, **41**, D348–D352.

Punta,M. *et al*. (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.

St Pierre,S.E. *et al*. (2014) FlyBase 102—advanced approaches to interrogating FlyBase. *Nucleic Acids Res.*, **42**, D780–D788.

Wernersso,R. and Pedersen,A.G. (2003) RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Res.*, **31**, 3537–3539.