# Remote Homology Detection Through Discriminative Statistical Relational Learning

Joel Bruno Santos da Costa[1], Juliana S Bernardes[1], Vítor Santos Costa[2], Gerson Zaverucha[1]

[1] Departament of Systems Engineering and Computer Science (COPPE), Federal University of Rio de Janeiro (UFRJ) - Rio de Janeiro - RJ, Brasil
{joel, julianab, gerson}@cos.ufrj.br

[2] Faculdade de Ciências, Universidade do Porto, Porto, Portugal
vsc@dcc.fc.up.pt

**Abstract.** An important problem in Computational Molecular Biology is the detection of remote homologues. We show that discriminative models, such as CRFs, can be useful in this task, and that we can elegantly encode structural information through logic, in frameworks such as TildeCRF.

**Keywords:** TildeCRF, Conditional Random Fields, Remote Homology Detection, HMMER.

## 1 Introduction

An important problem in Computational Molecular Biology is the detection of remote homologues, often proteins that have a common ancestor but that have diverged significantly in their evolutionary history. A number of tools have been developed toward this purpose. Arguably, a popular and effective approach is to model a family of proteins as a profile hidden Markov models (pHMM) [3]. Query proteins are aligned against the pHMM for the family. Such models are most often trained on primary sequence information only, although recent work has shown that greater sensitivity can be achieved by using secondary or tertiary information when available [2].

Hidden Markov models [10] provide a generative model for sequence data. Recently, there has been much interest in discriminative models of sequence data, such as conditional random fields (CRFs) [7]. In recent work, Kersting *el al.* [5] proposed TildeCRF, an extension to CRFs where the sequence is formed by logical atoms, thus providing a natural framework for expressing structured data. Initial LogCF results in predicting secondary structure were very promising.

In this work, we investigate whether probabilistic logical models of sequences can be competitive models for detecting remote homologues. In our experiments, we encode sequences with structural information as logical atoms, and we use the boosting based learning of TildeCRF to generate a discriminative model.

This paper is organized as follows. After discussing related work in Section 2, we will briefly review CRF and TildeCRF in the next two sections. We describe our

methodology and experiments in Section 5. Preliminary results are discussed in Section 6. Finally, Section 7 presents our conclusions and cites future work we intend to do.

## 2 Related work

Traditional approaches for homology detection are based on sequence information, i.e. they search the database using PSI-BLAST [13] or match against a profile hidden Markov model (pHMM) that describes preserved residues of protein sequences by creating a statistical model of aligned sequences [3]. These methods work well for simple conserved structures with strong sequence similarities, but fail for remote homology. More precisely, in the twilight zone of homology, where sequence similarity across proteins is poor, but structure is quite often well conserved, Bernardes *et al.* [2] show that alignments obtained using structure information can lead to more expressive models than alignments obtained from sequence data only.

This raises the question of how best one can improve sequence similarity detection through structural information. One approach relies on the widely available pHMM implementations, such as SAM [11] and HMMER [4], are widely available. Systems such as HMMER-STRUCT extend HMMER by considering structural information when computing the probabilistic model [9,27]. HMMER-STRUCT achieved good experimental results. Experiments showed that secondary and tertiary structural information could improve model quality, independently. In a related experiment, it was shown that there was a benefit in using packing and accessibility information, which is readily available.

HMMER and SAM are very efficient implementations of a generative propositional model for sequence data. Recently, there has been exciting work on discriminative models, such as CRFs [7], on the one hand, and on logical models of structured sequences, on the other hand. PRISM [24], SLPs[25], CLP(BN) [26], LoHMMs [6], Relational Markov networks (RMN) [15] , Relational Markov models [16] and TildeCRF are different approaches to this approach, but all can be used to models structured sequence data. PRISM and CLP(BN) have been used to implement HMMs. They provide very general frameworks, but allow more compact descriptions in the model. In contrast, LoHMMs is a framework specifically designed to handle sequences of logical atoms. As a very different alternative, Markov Fields have also been upgrade to first order logic, as in Models Relational Markov networks (RMN) [15] and Markov logic network (MLN) [14]. In the same spirit as LoHMM, TildeCRF can be seen as an attempt towards downgrading such highly expressive frameworks for the specific goal of handling logical sequences [5].

Interest in CRFs is further motivated by successful applications of CRFs in several bioinformatics applications, such as protein secondary structure prediction [5], protein fold classification [5,20] and RNA secondary structural alignments [21].

## 3 Conditional Random Fields

CRFs are undirected graphical models to compute a conditional probability distribution P(Y|X), where X is a input sequence that we assume has been observed, and Y is a set of output variables that we wish to predict. The principal advantage of discriminative modelling is that conditional distribution P(Y|X) does not include a model of P(X), which often contains many highly dependent features [17].

Formally, in special case of a linear chain structure, let G be an undirected graphical model over sets of random variables X and Y the labelling of an observed sequence X. Then a CRF defines the conditional probability as

$$P(Y|X) = Z(X)^{-1} \exp \{ \sum_{1..K} \lambda_k f_k(y_t, y_{t-1}, x_t) \} \quad (\mathbf{1})$$

where $\{f_k(y, y', x_t)\}_{k=1..K}$ is a set of real-valued feature functions, witch are given and fixed, Z(X) is a normalization factor and $\{\lambda_k\}_{k=1..K}$ is a parameter vector that will be learned via maximizing the conditional likelihood of the training data. In a linear chain CRF, a first order Markov assumption is made on the hidden variables and there is one feature per transition and one feature per state-observation pair, very much as in HMMs.

Parameter estimation is frequently performed by conditional log likelihood and optimized by gradient-based techniques. Because often we have a large number of parameters, a kind of penalty called *regularization* is adopted to avoid overfitting. Inference tasks can be performed efficiently by variants of the standard HMM algorithms, such as the Viterbi algorithm for finding the most likely explanation.

## 4 TildeCRF

In this work we rely on TildeCRF, to the best of our knowledge the first system that can train conditional random fields on *logical sequences*. The key idea of TildeCRF is to use relational regression trees in Dietterich *et al.*'s gradient tree boosting approach [19]. Following Dietterich's work, TildeCRF's potential functions are represented as weighted sums of regression trees. On the other hand, in TildeCRF regression trees are *relational*, as in Tilde [18]. Relational regression trees allow abstraction through logical variables and unification.

The compactness and even comprehensibility of TildeCRF, however, comes at the expense of a complex parameter estimation problem: the system relies on a non-parametric functional representation. Therefore, gradient-based optimization techniques such as McCallum's MALLET [23], which assume a parameterized representation, cannot be applied. Instead, TildeCRF follows Dietterich *et al.*'s gradient tree boosting technique [19], called TreeCRF. In TreeCRF, potential functions are represented by sums of traditional regression trees, which are grown stage-wise by a variant of boosting. Each regression tree can be viewed as defining several new feature combinations one corresponding to each path in the tree from the

root to a leaf. The resulting potential functions still have the form of a linear combination of features, but the features can be quite complex.

Boosting is implemented in a style similar to Dietterich *et al*.: one evaluates the gradient function at every position in every training example and fits a regression tree to these derived examples. But, simplify the derivation of the gradient and afterwards the evaluation, it does not use the complete input but a window. Relational regression trees upgrade the attribute value representation used within classical regression trees: every test is a relational conjunction of atoms.

In order to induce a relational regression tree, TildeCRF essentially employs Blockeel and De Raedt's Tilde, which also explains the name of its approach. Tilde learns relational trees by applying the *learning from interpretations* setting, where each example is an interpretation, or more precisely, a set of ground atoms. To learn, Tilde basically follows Quinlan's well-known C4.5 algorithm.

## 5 Methodology

In our study, we work at the super family level of the SCOP database [1], which groups families such that a common evolutionary origin is not obvious from sequence identity, but probable from an analysis of structure and from functional features. We believe that this level best represents remote homologies.

We aim at investigating the following open problems: **(i)** does TildeCRF achieve significantly better results by using structural information; **(ii)** is TildeCRF competitive with standard tools developed for this task, such as pHMMs.

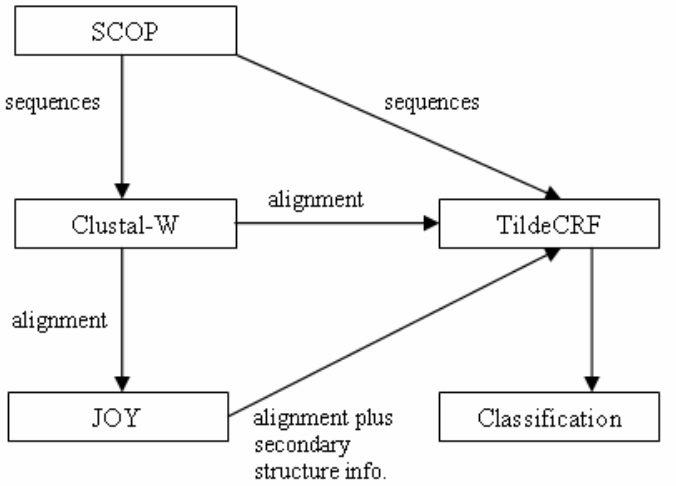


**Fig. 1.** Experiment schema: the information needed in each test is represented for the three arrows ending in TildeCRF box.

In a preliminary test, we select three super families of the alpha class. We then use cross-validation [8], leaving one-family out of the training, to evaluate our results. We repeat the process three times. In the first experiment, we use sequences of amino acids, thus including no structural information whatsoever. In a second experiment, we first align a new sequence against a multiple alignment for the family, obtained from Clustal-W [12]. In the third step we use information of the secondary structure of the protein, obtained using JOY [22]. Finally we compare these results with HMMER and HMMER-STRUCT results.

We have defined a simple form to represent the amino acids of the protein sequences into first-order logical sentences: a unique predicate "a" having an arity 1, 2 or 3, depending on the test being performed. The first term represents the amino acid properly and may receive one of 20 letters that are used to represent amino acids in biology. The second encode the secondary structure of protein witch that amino acid is part and receive "C" for coil, "H" for helixes, "E" for beta and "P" for phi angle. The last encode the alignment information, meaning "m" for matches, "i" for inserts and "g" for gaps. Then the input sequences are made up of sequences of amino acids represented like below.

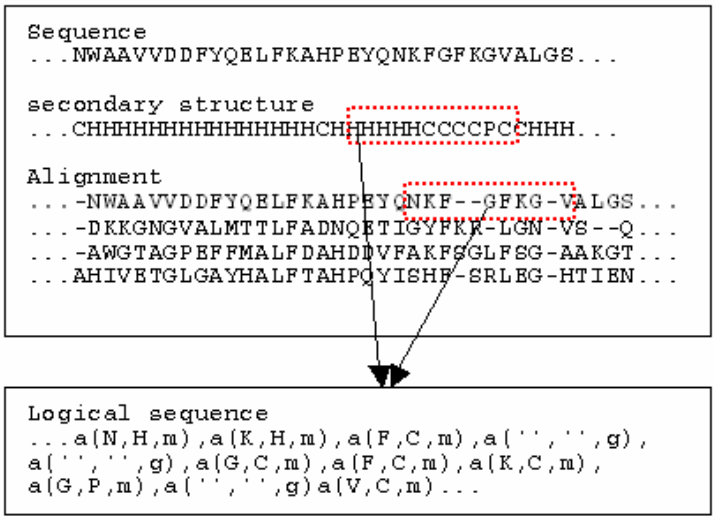


**Fig. 2.** Input sequence schema.

## 6 Evaluation

Three super families were chosen at random from SCOP alpha class: a.1.1, a.3.1 e a.4.1. Accuracy on training and test sets of Java version of TildeCRF are shown in

tables 1 e 2. The input features consisted of an 11-residue slide window and we allowed regression trees of depth 5 at maximum.

**Table 1.** Accuracy on training set

| Super family | Amino acid info. | Alignment info. | Secondary structure info. |
|---|---|---|---|
| a.1.1 | 0.43 | 0.87 | 0.94 |
| a.3.1 | 0.41 | 0.67 | 0.75 |
| a.4.1 | 0.31 | 0.71 | 0.80 |

Our results show that including more information usually improves both train-set and test-set accuracy. This affirmatively answers **(i)**.

**Table 2.** Accuracy on test set

| Super family | Amino acid info. | Alignment info. | Secondary structure info. |
|---|---|---|---|
| a.1.1 | 0.17 | 0.38 | 0.78 |
| a.3.1 | 0.20 | 0.37 | 0.34 |
| a.4.1 | 0.31 | 0.46 | 0.62 |

In order to assess the significance of the results was used *paired t-test* [8] considering the results as significant at $p = 0.05$.

**Table 3.** Paired t-test over test results (table 2): P value and statistical significance

| | Amino acid info. | Secondary structure info. |
|---|---|---|
| Amino acid info. | | 0.03934 (yes) |
| Alignment info. | 0.00277 (yes) | 0.00006 (yes) |

We next repeated the same experiments using HMMER, arguably one of the most popular tools in searching for remote homologues, and HMMER-STRUCT that build five pHMMs from the same train set, one for each structural property. The properties used are: primary, secondary and tertiary structures, accessibility and packing residue. Here we have used HMMER-STRUCT by considering only primary and secondary structure properties. Table 4 show results per super family.

**Table 4.** Comparing TildeCRF with others, accuracy on test set

| Super family | HMMER | HMMER-STRUCT<br>Secondary structure info. | TildeCRF<br>Secondary structure info. |
|---|---|---|---|
| a.1.1 | 0.67 | 0,74 | 0.78 |
| a.3.1 | 0,97 | 0,97 | 0.34 |
| a.4.1 | 0,54 | 0,56 | 0.62 |

HMMER achieved an average accuracy of 0.73 and HMMER-STRUCT 0.76. In contrast, TildeCRF achieves only 0.58. This results apparently shows that TildeCRF does not perform as well as the others. But it obtained best the results in two of three super families we considered, suggesting that question (ii) needs further research in improving the performance of TildeCRF through better usage of the structural information used by HMMER-STRUCT.

## 7 Conclusions and Future Work

We believe that discriminative statistical relational models, such as conditional random fields, can be beneficial in the important problem of remote homology detection.

As a next step, we will include further experiments. We are running the same experiments presented in this work for super families of SCOP beta and alpha-beta classes. Our intention is to investigate how TildeCRF performs in the three major classes of the SCOP database. After, we will enlarge the sampling, running it for as many super families as possible.

## References


1. Andreeva, A., Howorth, D., Brenner, S., Hubbard, T., Chothia, C., Murzin, A.: SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Research, v. 32, n. 1, pp. 226-229 (2004)
2. Bernardes, J. S., Davila, A. M. R., Santos Costa, V., Zaverucha, G.: Improving Model Construction of Profile HMMs for Remote Homology Detection Through Structural Alignment. BMC Bioinformatics, v. 8:435, p. 1-12 (2007)
3. Durbin, R., Eddy, S., Krogh, A., Mitchison, G.: Biological SequenceAnalysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge UK (1998)
4. Eddy, S.: Profile hidden Markov models. Bioinformatics, v. 14, n. 9, pp. 755-763 (1998)
5. Gutmann, B., Kersting, K.: TildeCRF: Conditional Random Fields for Logical Sequences. In Proc. of the 15th European Conf. on Machine Learning (ECML). Lecture Notes of Artificial Intelligence, v. 4212, pp. 174-185 (2006)
6. Kersting, K., De Raedt, L., Raiko, T.: Logical Hidden Markov Models. Journal of Artificial Intelligence Research, Volume 25, pages 425-456 (2006)
7. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. Proc. of 18th International Conf. on Machine Learning (ICML), pp. 282-289 (2001)
8. Mitchell, T. M.: Machine Learning. McGraw-Hill, New York (1997)
9. Bernardes, J. S.: Remote Homology Detection with HMM and Structural Issues. M.Sc. thesis, In Portuguese, Federal University of Rio de Janeiro, COPPE/UFRJ (2006)
10. Rabiner, L.: A tutorial on hidden Markov models and selected applications in speech recognition. In Proceedings of IEEE, volume 77, pp. 267-296 (1989)
11. Hughey, R., Krogh, A.: SAM: Sequence alignment and modeling software system. Technical Report UCSC-CRL-95-7, University of California, Santa Cruz (1995)

12.Thompson, J., Gibson, T.: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Computer Applications in the Biosciences, v. 22, n. 22, pp. 4673-4680 (1994)
13.Altschul, S. F., Madden T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J.: Gapped BLAST and PSI-blast: a new generation of protein database search programs. Nucleic Acids Res., 25(17): 389–402, (1997)
14.Richardson, M., Domingos, P.: Markov Logic Networks. Machine Learning, 62:107-997 (2006)
15.Taskar, B., Abbeel, P., Koller, D.: Discriminative Probabilistic Models for Relational Data. In Proc. of the 8th Conf. on Uncertainty in Artificial Intelligence(UAI-02), pp. 485-492, (2002)
16.Anderson, C. R., Domingos, P., Weld, D. S.: Relational Markov Models and their Application to Adaptive Web Navigation. In Proc. of the 8th Int. Conf. on Knowledge Discovery and Data Mining (KDD-02), pp. 143-152 (2002)
17.Sutton, C., McCallum, A.: An Introduction to Conditional Random Fields for Relational Learning. Book chapter in Introduction to Statistical Relational Learning. Edited by Lise Getoor and Ben Taskar. MIT Press. (2006)
18.Blockeel, H., De Raedt, L.: Top-down Induction of First-order Logical Decision Trees. Artificial Intelligence, 101(1-2):285-297 (1998)
19.Dietterich, T., Ashenfelter, A., Bulatov, Y.: Training conditional random fields via gradient tree boosting. In Proc. 21th International Conf. on Machine Learning, pp. 217-224. ACM (2004)
20.Liu, Y., Carbonell, J., Weigele, P., and Gopalakrishnan, V.: Segmentation conditional random fields (SCRFs): A new approach for protein fold recognition. In ACM International conference on Research in Computational Molecular Biology - RECOMB05 (2005)
21.Sato, K., Sakakibara, Y.: RNA secondary structural alignment with conditional random fields. Bioinformatics, 21:ii237-242 (2005).
22.Mizuguchi, K., Deane, C., Blundell, T., Johnson, M., Overington, J.: JOY: protein sequence-structure representation and analysis, Bioinformatics, v. 14, n. 7, pp. 617-623 (1998)
23.McCallum, A.: Efficiently inducing features of conditional random Fields. In Proc. of the 21st Conference on Uncertainty in Artificial Intelligence (2003).
24.Sato, T., Kameya, Y.: PRISM: A symbolic-statistical modeling language. Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI), pp.1330–1335 (1997)
25.Muggleton, S.: Stochastic logic programs. In De Raedt, L., ed., Advances in Inductive Logic Programming, 254–264. IOS Press (1996)
26.Santos Costa V., Page, D., Cussens J: CLP(BN): Constraint Logic Programming for Probabilistic Knowledge. Probabilistic Inductive Logic Programming, pp. 156-188 (2008)
27.Bernardes, J. S., Davila, A. M. R., Santos Costa, V., Zaverucha, G.: HMMER-STRUCT: Adding structural properties to profiles HMMs. (To be submitted).