



Improved smoothing for Probabilistic Suffix Trees seen as variable order Markov chains

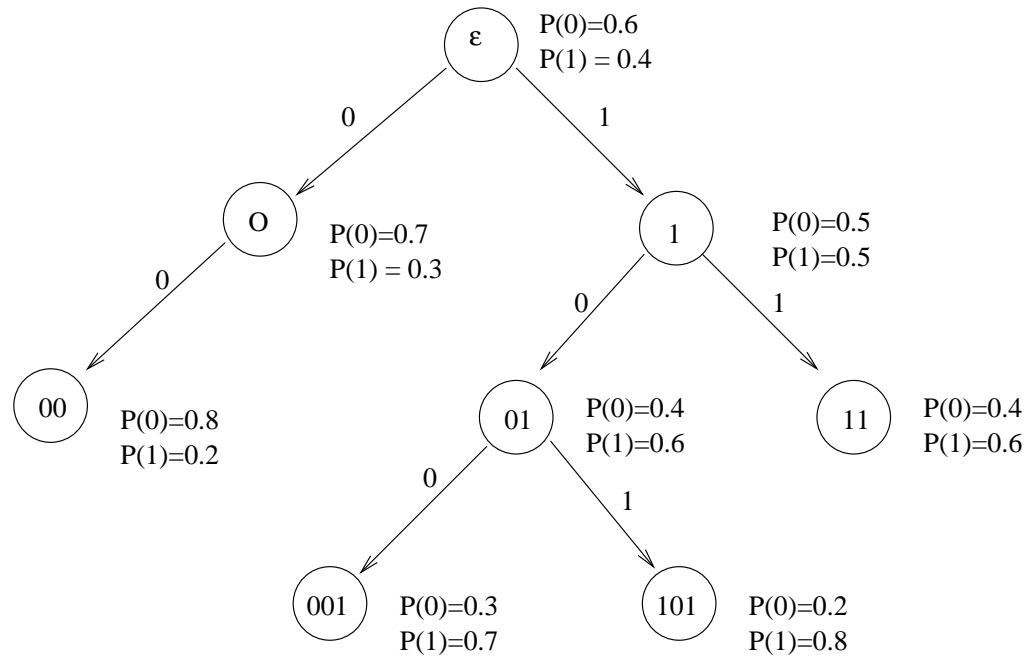
Christopher Kermorvant¹ and Pierre Dupont²

¹EURISE, Université Jean Monnet, Saint-Etienne, France

²INGI, University of Louvain, Louvain-la-Neuve, Belgium

Probabilistic Suffix Tree

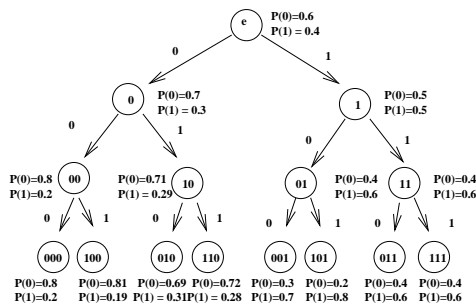
- PST assign probabilities to sequences
- a PST models a set of conditional distributions
 $P(\sigma|h)$: probability of a symbol σ given a suffix h , denoted $\gamma_h(\sigma)$
- the probability of a sequence according to a PST is the product of the probability of each letter given its longest suffix in the PST.



$$\begin{aligned}
 P(100110) &= \gamma_e(1)\gamma_1(0)\gamma_0(0)\gamma_{00}(1)\gamma_{001}(1)\gamma_{11}(0) \\
 &= 0.4 * 0.5 * 0.7 * 0.2 * 0.7 * 0.4 \\
 &= 7.84 \cdot 10^{-3}
 \end{aligned}$$

PST / Markov Chain relationship

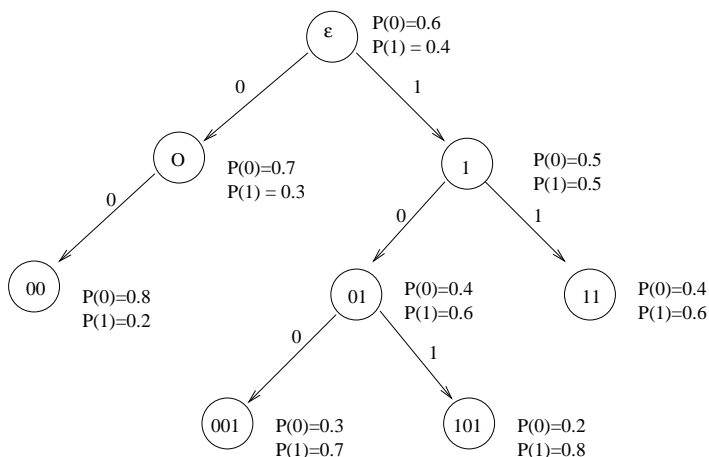
- complete PST of depth L = Markov chain of order L
- non complete PST = variable order Markov chain (the order depends on the suffix)



$$\Leftrightarrow P(s_n | s_1, \dots, s_{n-1}) = P(s_n | s_{n-3} s_{n-2} s_{n-1})$$

Problems with PST

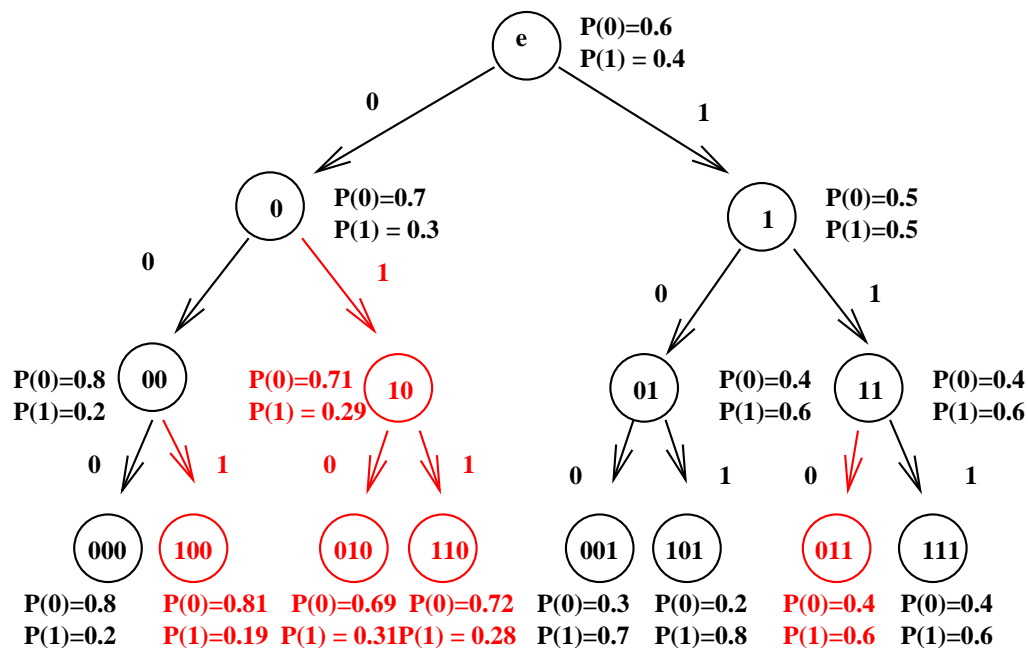
1. the size of the complete tree is exponential in the maximal size of suffixes (depth of the tree)
2. “zero-frequency problem” : estimation on limited data, some probabilities are badly estimated (need for smoothing)



$$P(1|10) = ?$$

Solution for the size

Prune the tree : keep a node only if its distributions are different enough from its parent's distributions :



Solutions for smoothing

- basic solution : add a flooring value
- better solution : Back-off smoothing
 - discount a probability mass from seen events
 - redistribute it on unseen events according to a back-off (e.g. lower order) distribution

The Back-off smoothing of a Markov chain defines a variable order Markov chain

Smoothing

$$P(\sigma|s) = \begin{cases} \frac{c(s,\sigma) - d_C}{\sum_{\sigma \in \Sigma} c(s,\sigma)} & \text{if } c(s,\sigma) > 0 \\ \alpha(s)\beta(s,\sigma) & \text{otherwise} \end{cases}$$

- $c(s, \sigma)$: number of times σ was seen after the suffix s
- d_C is the discount parameter
- $\alpha(s)$ is a normalization factor
- $\beta(s, \sigma)$ is the back-off distribution

Smoothing

Even better : non-shadowing scheme
Use both the main and back-off distributions
whenever possible

$$P(\sigma|s) = \begin{cases} \frac{c(s,\sigma)-d_C}{\sum_{\sigma \in \Sigma} c(s,\sigma)} + \alpha'(s)\beta(s,\sigma) & \text{if } c(s,\sigma) > 0 \\ \alpha'(s)\beta(s,\sigma) & \text{otherwise} \end{cases}$$

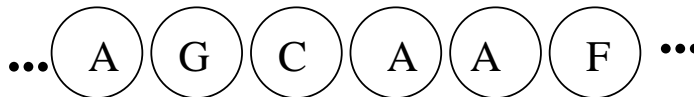
Application : Protein domain detection

- Proteins, main components of living beings : metabolism, cell structure, energy, defense.
- Sequence of amino-acids : 20 letters alphabet

Protein structure

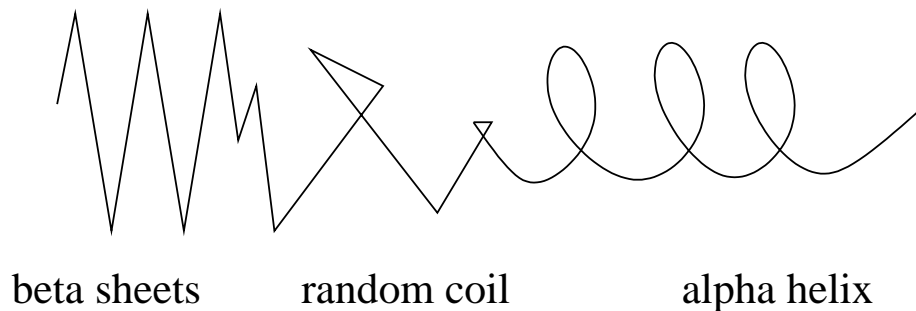
Primary Structure

amino-acids sequence

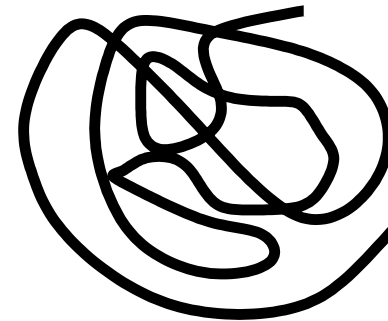


Secondary structure

regular or irregular



Tertiary structure



Quaternary structure

Association of several proteins

Domains and functions :

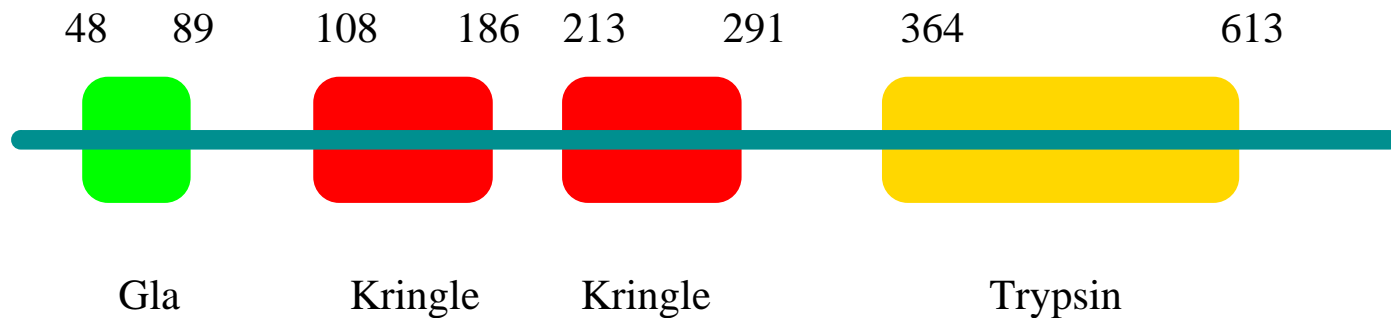
- Domain : subsequence of a protein
- Domains are functional units.
- 1 or several different domains in a protein

Sequence homology \Rightarrow structure and function similarity

Domains detection in an unknown protein allows to make hypothesis about its functions or structure

Domain detection:

THRB_HUMAN (Thrombine)



Unknown protein



What are the domains ? Where are they ?

The problem : domain detection

Experimental setup :

[Bejerano01][Eskin00]

- Proteins from SWISSPROT
- Domains from PFAM
- Evaluation criterion : based on iso-point (same number of false + and false -)

Databases

- SWISSPROT *Proteins* database :
 - sequences + annotations (functions, structures, domains, etc.)
 - SPROT 33 (02/1996) : 52 205 sequences, 18 531 384 AA
- PFAM *Domains* database :
 - sequences + annotations (functions, proteins)
 - PFAM 1.0 (04/1996) : 175 families, 15 610 sequences, 3 560 959 AA.

Experimental setup :

- train PST and Markov Chains on 80% of PFAM
- test on 100% SWISSPROT
- compute a family threshold per family s.t. same number of false + and false - (iso-point)
- compute % true positives $>$ iso-point

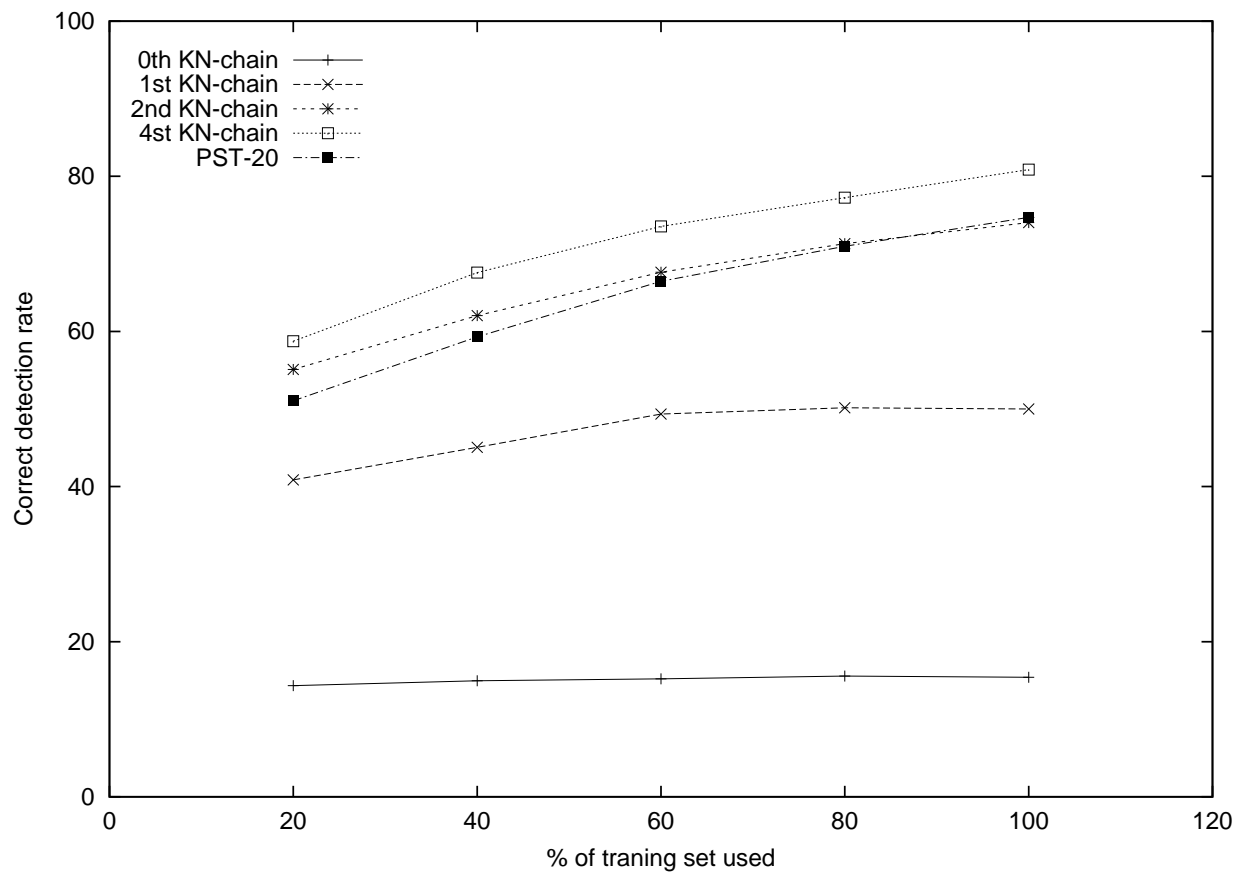
Results

Correct detection rate (SWISSPROT database) :

	Non-shadowing (KN) Smoothing					PST
Maximal order	0	1	2	3	4	20
Correct detection rate	13.9	53.0	81.3	89.5	90.0	85.8
Number of parameters	$2,2 \times 10^1$	$4,0 \times 10^2$	$3,3 \times 10^3$	$9,9 \times 10^3$	$1,8 \times 10^4$	$5,1 \times 10^4$

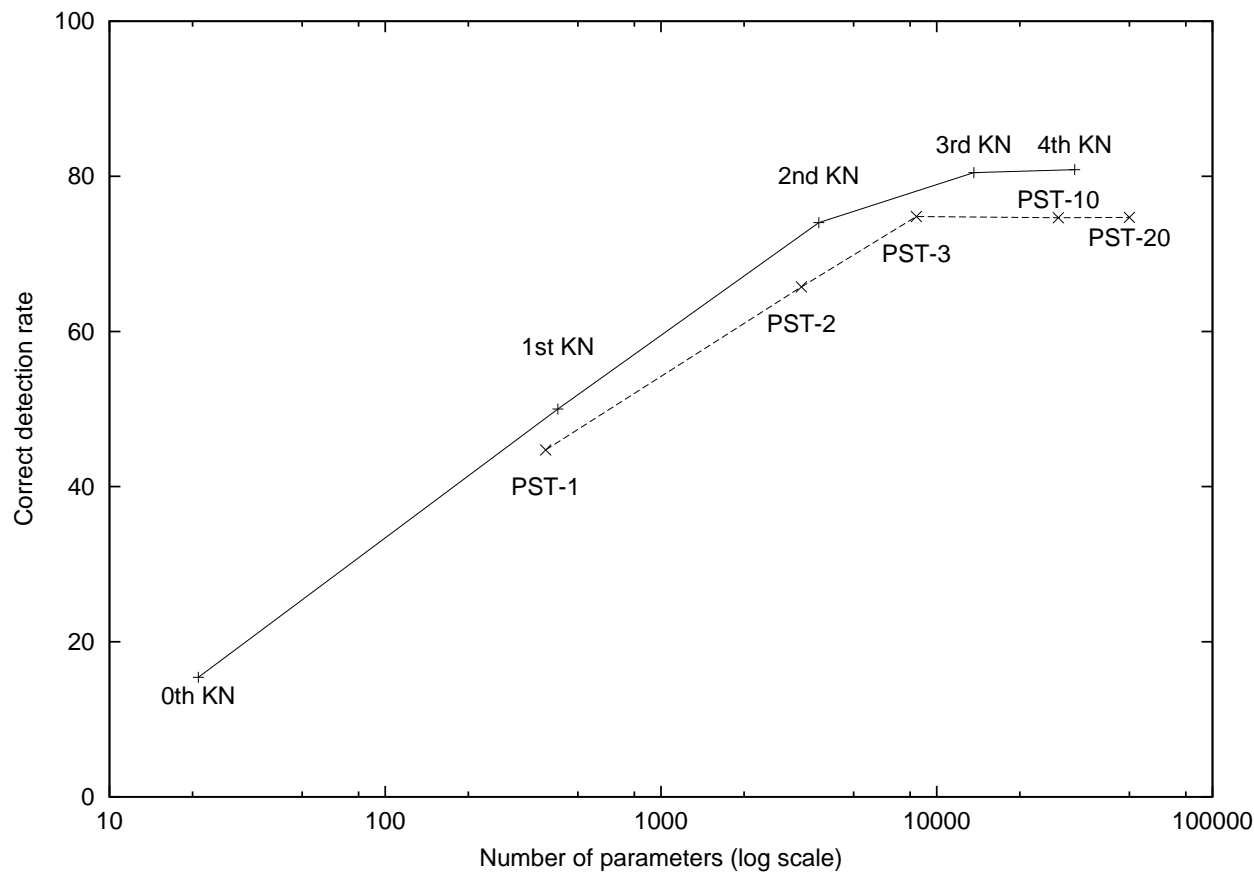
Results

Correct detection rate vs size of the training set :



Results

Correct detection rate vs number of parameters :



Conclusion

Summary of the best strategy :

- no pruning
- use both the main and back-off distributions whenever possible
- better correct detection rate with fewer parameters

Future directions

- make a detailed comparison with HMM
- investigate a small sample framework
 - better statistical test for pruning trees
 - statistical distances between trees