ppALIGN: Posterior distribution for score-based alignments

Stefan Wolfsheimer
Gregory Nuel

Mathématiques Appliqués à Paris 5
Université Paris Descartes

Outline

Score-based alignment

Probabilstic alignment: ppALIGN

ppALIGN in action

What is an alignment ?

Example (Two DNA sequences)

$$a_1^\ell = \text{a c g t a g c a t g a c a}$$
$$b_1^m = \text{a c c g t a c a a g c a}$$

c: common ancestral sequence

c	a	c	c	g	t	t	a		c	a	a	g	a	c	a
a_1^ℓ	a	c		g	t		a	g	c	a	t	g	a	c	a
b_1^m	a	c	c	g	t		a		c	a	a	g		c	a

here is the alignment $\mathcal{A} = (\widetilde{a}_1^t, \widetilde{b}_1^t)$ we get:

$\widetilde{a}_1^t$	a	c	-	g	t		a	g	c	a	t	g	a	c	a
$\widetilde{b}_1^t$	a	c	c	g	t		a	-	c	a	a	g	-	c	a

Note that $\max(\ell, m) \leq t \leq \ell + m$.

Score-based alignments

Definition (Score of an Alignment)

- scoring function: $\sigma : \Sigma \cup \{\text{-}\} \times \Sigma \cup \{\text{-}\} \to \mathbb{R}$
- score of an alignment: $s(\mathcal{A}) = s(\widetilde{a}_1^t, \widetilde{b}_1^t) = \sum_{k=1}^{t} \sigma(\widetilde{a}_k, \widetilde{b}_k)$

- optimal score $s_0 = \max_{\mathcal{A}} s(\mathcal{A}; a_1^\ell, b_1^m)$ and alignment $\mathcal{A}_0 = \operatorname{argmax}_{\mathcal{A}} s(\mathcal{A}; a_1^\ell, b_1^m)$
- Needleman-Wunsch algorithm (1970)
- Problem in score-based alignment: (nearly) optimal alignments not unique.
- Reliable regions: common to all high scoring alignments
- Questionable regions: close to gaps and low complexity regions

Posterior probabilities

Example (protein alignment)

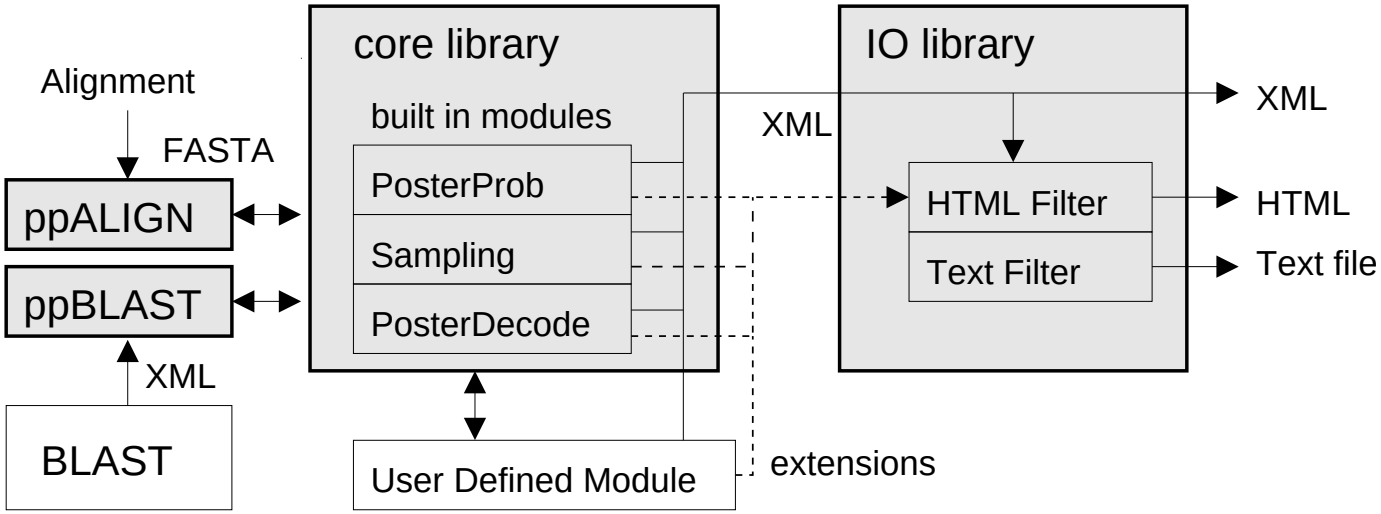
S A L L A S G G T S S H R W S R T S A L L M A R K S H R V L W S R T	score = 31
S A L L A S G G T S S H R - - W S R T S A L L M A - - R K S H R V L W S R T	score = 31
S A L L A S G G T S S H R - - W S R T S A L L - - M A R K S H R V L W S R T	score = 28

score = 31

score = 31

score = 28

ppALIGN



Features

- Alternative decoding algorithms
- Input filter for BLAST
- Structured output (XML)
- Standalone programs and C++ library (open source)
- Webinterface http://www.math-info.univ-paris5.fr/ppblast/

SW, AK Hartmann, G Nuel, preprint

ppALIGN • 1.0
posterior probabilities for
score based sequence alignments

Back | Protein Alignment | DNA Alignment

Protein alignment

SCORE PARAMETERS

Score Matrix: ? blosum62

Gap open: ? 11

Gap extension: ? 1

ppALIGN PARAMETERS

Model: ? (•) Pair Hidden Markov Model

? ( ) Finite temperature alignment

ALGORITHM

Algorithm: ? (•) Global alignment

? ( ) Local alignment

? ( ) Boundaries of local alignments

Alignment: ? ( ) Find optimal alignment

? (•) User defined alignment

ALTERNATIVE ALIGNMENTS

Decoding method: ? [☑] Maximum accuracy alignment

? [☑] Sampling from the posterior distribution

Number of sampels: 10

USER DEFINED ALIGNMENT

Compute the posterior probabilities of the following global alignment.
The number of columns in both sequences must agree.

Alignment (Fasta Format):

```
>62810|unnamed protein product
LQKVSTPQPVNAGS--EDESGKG------NLGFIHAFVASISVIIVSELGDKTFFIAAIM
AMRYNRLVVLAGAMLALGVMTCLSVLFGYATTIIPRIYTYYVSTALFAIFGIRMLREGLK
>gi|170036348|ref|XP_001846026.1|conserved hypothetical protein [Culex
quinquefasciatus]
VTELSNPNVESSGSPGEKKSSAGGGLLSSDVGFMHAFIASFSVIIVSELGDKTFFIAAIM
AMRHPRLTVFAGAIAALALMTVLSAVFGMAATIIPRVYTYYISTALFALFGLKMLKEGYY
```

ppALIGN

ppALIGN Result

Parameters

[Show]

Alignment

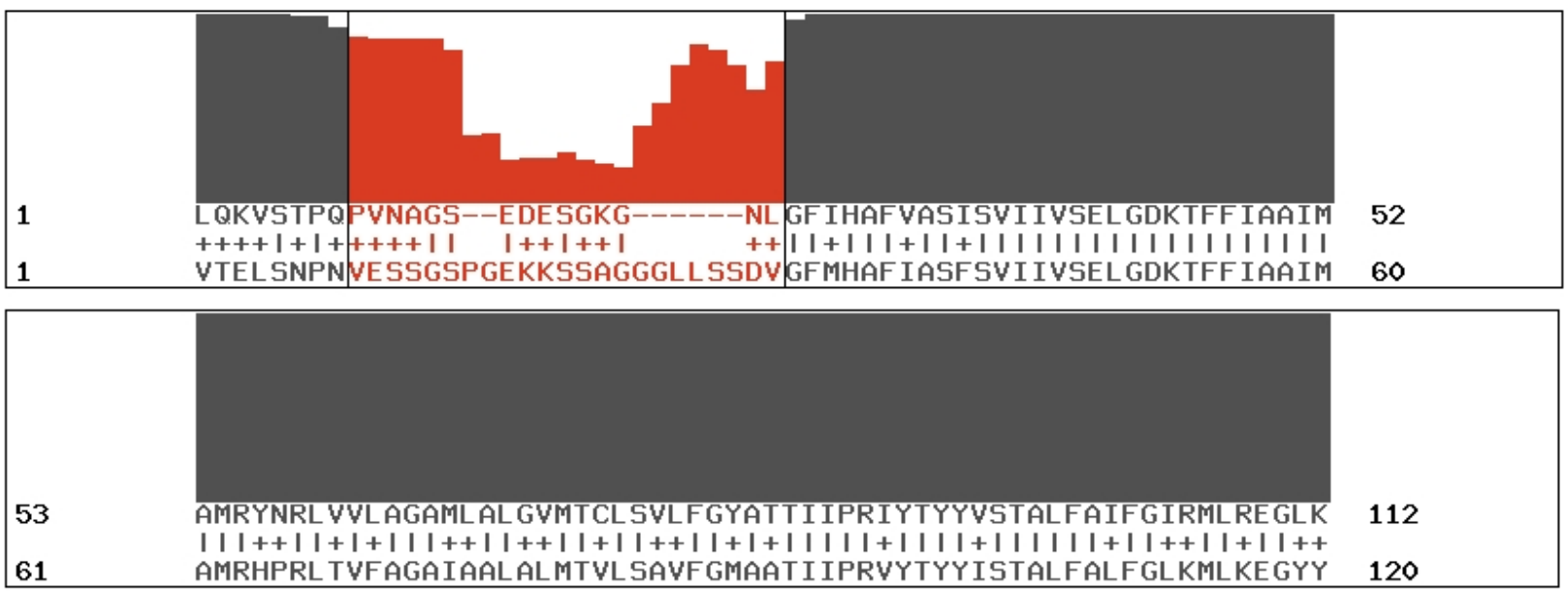
[Show Details]

Regions common to all alignments (your provided one and alternative alignments)

Regions not common to all alignments. Click on those regions to display possible alternatives

maximum marginalized averaged posterior

Sampled from the posterior distribution.



ppALIGN Result

Parameters

[Show]

Alignment

[Show Details]