Stochastic Pairwise Alignmen

Ulrike Mückstein

Institute for Theoretical Chemistry and Structur

University Vienna

http://www.tbi.univie.ac.at/~ulim/

Bled, 2003

Motivation

- ⋆ Alignment algorithms normally return a single optimal a optimal alignment does not need to be the only optir the optimization problem. Additionally suboptimal alignr equal quality might exist.

- ⋆ The optimal alignment of distantly related sequences ma ceptible to small perturbations of scoring parameters.

- ⋆ The reliability of an alignment depends strongly on th sequence similarity.

Approaches dealing with locally variable alignmen

Variations of dynamic programming algorithms to constr
alignments
Vingron & Argos, 1990;
Saqi & Sternberg, 1991;

Calculation of the *partition function* and of match proba
Miyazawa, 1994;

Generation of ensembles of suboptimal alignments by s
tracking
this work;

Sequence Alignments

score $s(a, b)$ of a match between two residues

$$s(a, b) = k \log \frac{f_{ab}}{f_a f_b}$$

additive alignment score function $S(\mathcal{A})$

$$S(\mathcal{A}) = S(\mathcal{A}_{1,1}^{i-1,j-1}) + s(a_i, b_j) + S(\mathcal{A}_{i+1,j-$$

Probabilistic Interpretation of Sequence A

probability of a particular alignment $\mathcal{A}$

$$\text{Prob}(\mathcal{A}) = \frac{1}{Z} \exp\left(\frac{S(\mathcal{A})}{k}\right)$$

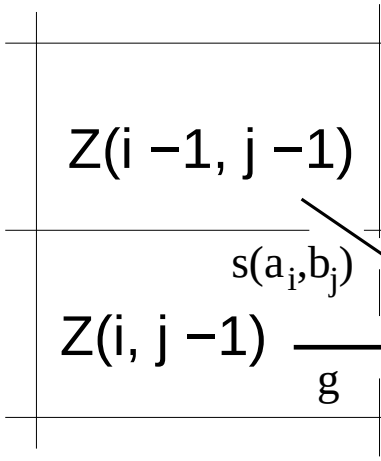
partition function

$$Z = \sum_{\mathcal{A}} e^{\frac{S(\mathcal{A})}{k}}$$

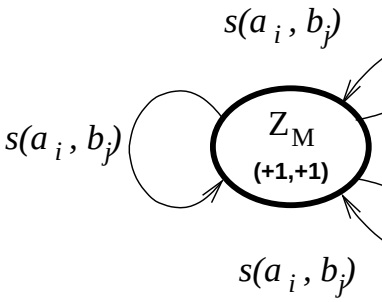
see e.g. Miyazawa, 1994, Yu & Hwa, 2001.

Partition Function

$$
\begin{aligned}
Z_{i,j} &= Z_{i-1,j-1} e^{\beta s(a_i, b_i)} \\
&\quad + Z_{i,j-1} e^{\beta g} \\
&\quad + Z_{i-1,j} e^{\beta g}
\end{aligned}
$$



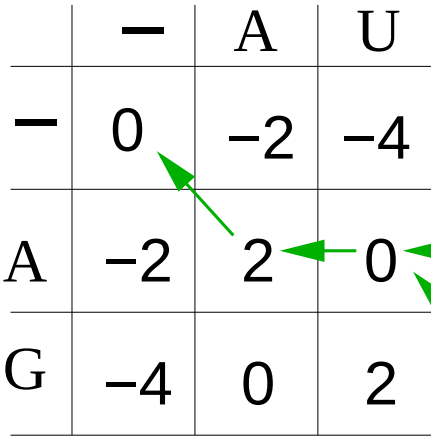
$\beta = \frac{1}{k}$



canonical Alignments

```
A---XXXXB   and   AXXXX---B
AYYY----B         A----YYYB
```

Stochastic Backtracking



next state	next indices $i \leftarrow$	$j \leftarrow$	transition probability
match	$i-1$	$j-1$	$\frac{Z_{i-1,j-1}}{Z_{i,j}} e^{\beta s(a_i, b_j)}$
gap in $\mathbf{a}$	i	$j-1$	$\frac{Z_{i,j-1}}{Z_{i,j}} e^{\beta g}$
gap in $\mathbf{b}$	$i-1$	j	$\frac{Z_{i-1,j}}{Z_{i,j}} e^{\beta g}$

Match Probabilities

$$P_{ij} = \frac{Z^{\mathsf{M}}_{ij} \widehat{Z}^{\mathsf{M}}_{ij}}{Z} \exp(-\beta s(a_i, b_j)$$

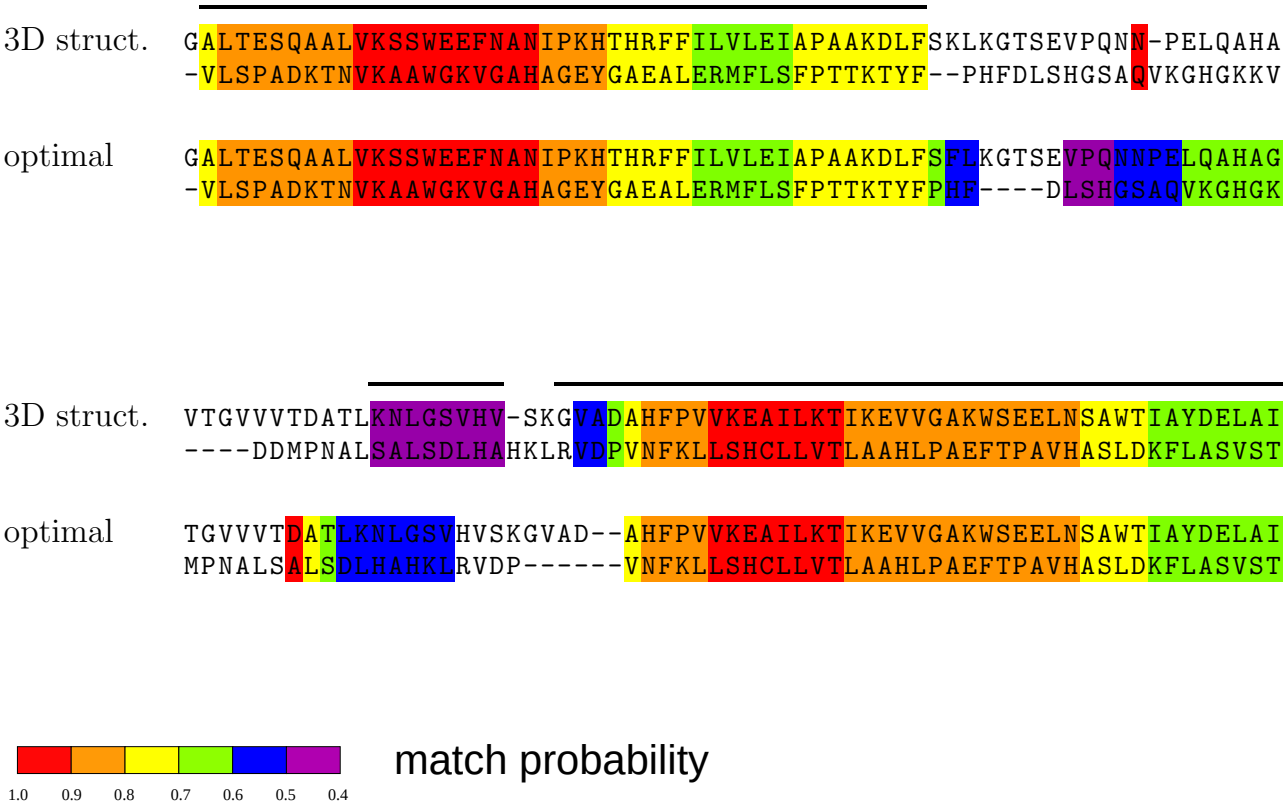
partition function of all alignments

Z^{M}_{ij} of the sub-sequences $\mathbf{a}[1..i]$ and $\mathbf{b}[1..j]$ ending with a match of (a_i, b_j)

$\widehat{Z}^{\mathsf{M}}_{ij}$ of the sub-sequences $\mathbf{a}[i..m]$ and $\mathbf{b}[j..n]$ starting with a match of $(a_i, b$

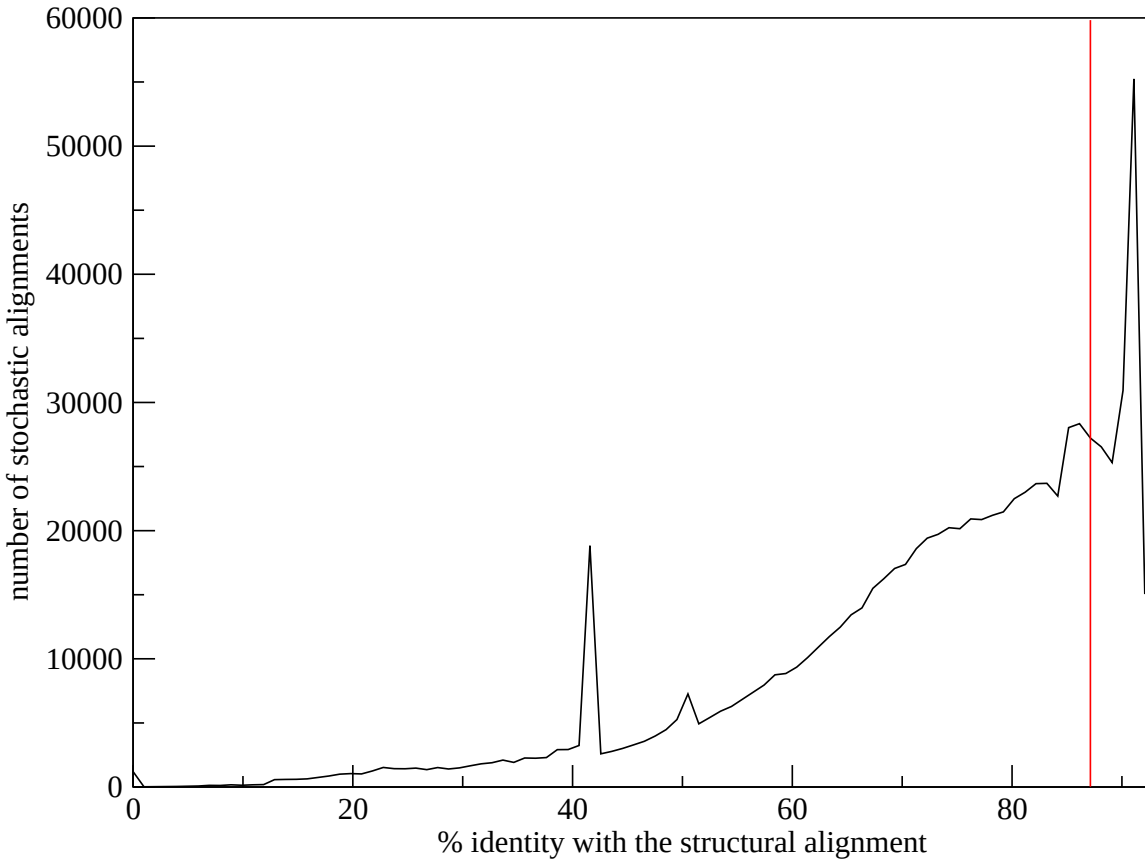
see Miyazawa, 1994, Vingron & Argos, 1990.

Comparison with Structure Based Alignme



alignment of *L.luteus* Leghaemoglobin / human Hemoglobin.

Stochastic Backtracking Finds Suboptimal Alignments that a
the Optimal Alignment

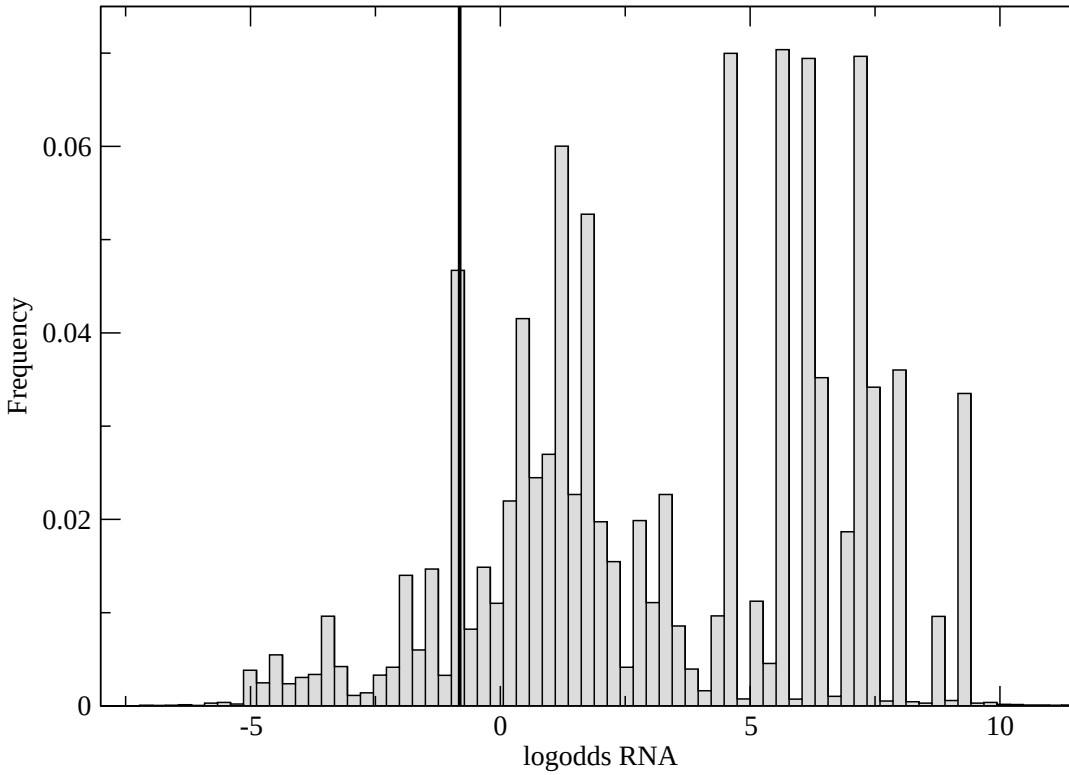


Comparison with Structural Alignments by CE

PDB entry	length	% identity		% eq
		optimal	mean of stoch. aln.	be than
1YCC/1CTJ[1]	108/89	51.3	50.3	4
4HHB/1GDJ[2]	141/153	67.2	63.5	3
1SAC/1C4R[3]	204/182	44.4	38.4	2

[1] Cytochrome C, *S.cer.* / Cytochrome C6, *M.braunii*
[2] Hemoglobin, Human / Leghaemoglobin, *L.luteus*
[3] Serum Amyloid P Component, Human / Neurexin (partial), *R.nor.*

Using Stochastic Backtracking with other M



Classification of the TAR hairpin of HIV with `qrna`, Rivas & Ed

Conclusion

- In this work we present an algorithm that produces cor
  samples of alignments by stochastic backtracking.

- The ensemble of stochastic alignments contains correct a
  significant probabilities even though the optimal alignme
  nificantly from the structural alignment.

- Stochastic pairwise alignments can be used as input data
  formatic tools.

- The software package can be downloaded from the inter
  `http://www.tbi.univie.ac.at/~ulim/probA/`.

- iterative multiple alignment procedures are likely to be tra
  pairwise alignments that differ from the true alignment

- use the match probability matrix of the pairwise alignmer
  of stochastic pairwise alignments to develop a multiple a

do not use canonical alignments

```
A---XXXXB   and   AXXXX---B
AYYY----B         A----YYYB
CXXXXYYYC         CXXXXYYYC
```

Thanks to

Peter Stadler, Ivo Hofacker, Peter Schuster.