A Variation on Algorithms for Pairwise Global Alignments

Ulrike Mueckstein

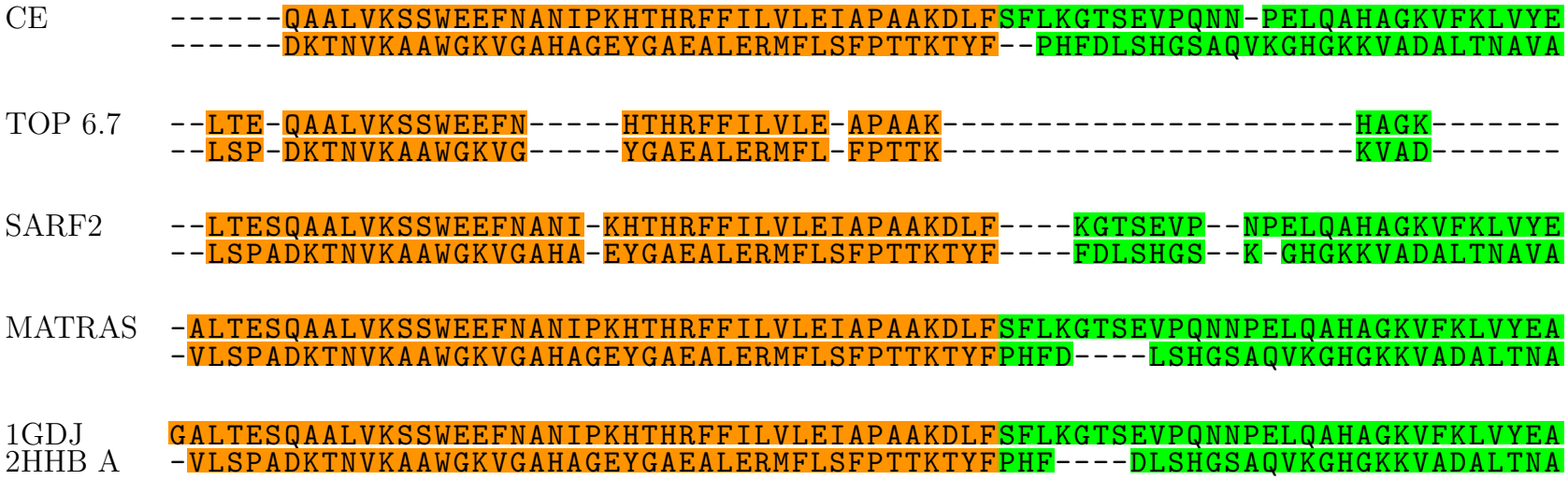
Institute for Theoretical Chemistry and Structural Biology

University Vienna

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

Bled, 2002

Various Alignments



alignment reliability

high	low

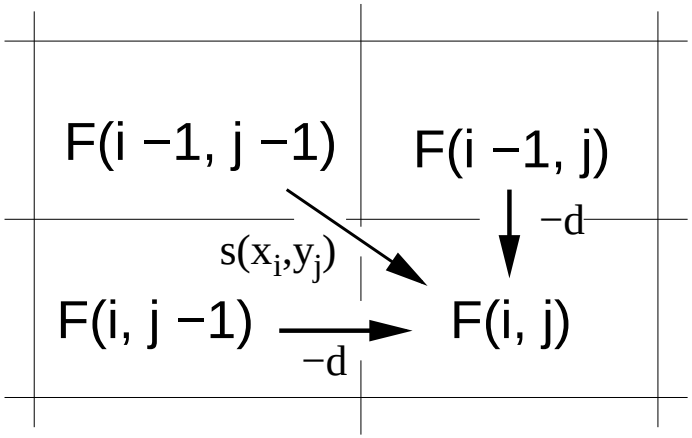
alignment between *Lupinus luteus* leghaemoglobin (1GDJ) and human chain A deoxyhaemoglobin (2HHB A)

Needleman-Wunsch Algorithm

★ dynamic programming algorithm: an optimal alignment is build recursively, using previous solutions for optimal alignments for smaller subsequences.

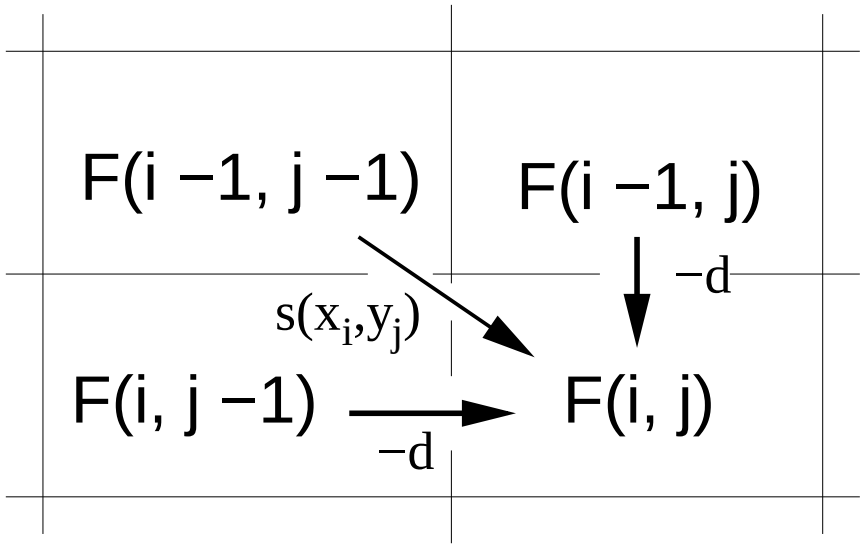
$XXXx_i$	$XXXx_i$	XXx_i-
$YYYy_j$	YYy_j-	$YYYy_j$

★ The score of the best alignment between the initial segments $x_1, x_2, ..., x_i$ and $y_1, y_2, ..., y_j$ is stored in matrix F at position F(i, j).



Recursion

- match (x_i, y_j):
  $\mathsf{F(i,j) = F(i-1,j-1) + s(x_i, y_i)}$
- gap in sequence **x** $(-, y_j)$:
  $\mathsf{F(i,j) = F(i,j-1) - d}$
- gap in sequence **y** $(x_i, -)$:
  $\mathsf{F(i,j) = F(i-1,j) - d}$



$$F(i,j) = \max \begin{cases} F(i-1,j-1) & + & s(x_i, y_i) \\ F(i-1,j) & - & d \\ F(i,j-1) & - & d \end{cases}$$

★ Initialisation

$F(0,0) = 0; \quad F(i,0) = -id; \quad F(0,j) = -jd;$

★ Extension

$$F(i,j) = max \begin{cases} F(i-1,j-1) & + & s(x_i,y_i) \\ F(i-1,j) & - & d \\ F(i,j-1) & - & d \end{cases}$$

★ Backtracking

F(n,m), is by definition the best score for an alignment of $x_1, x_2, ..., x_n$ and $y_1, y_2, ..., y_m$.

	–	A	U	G	G
–	0	–2	–4	–6	–8
A	–2	2	0	–2	–4
G	–4	0	2	2	0

$s(x,y) = 2$ if $x = y$;
$s(x,y) = 0$ if $x \neq y; d = 2$

Gap penalties

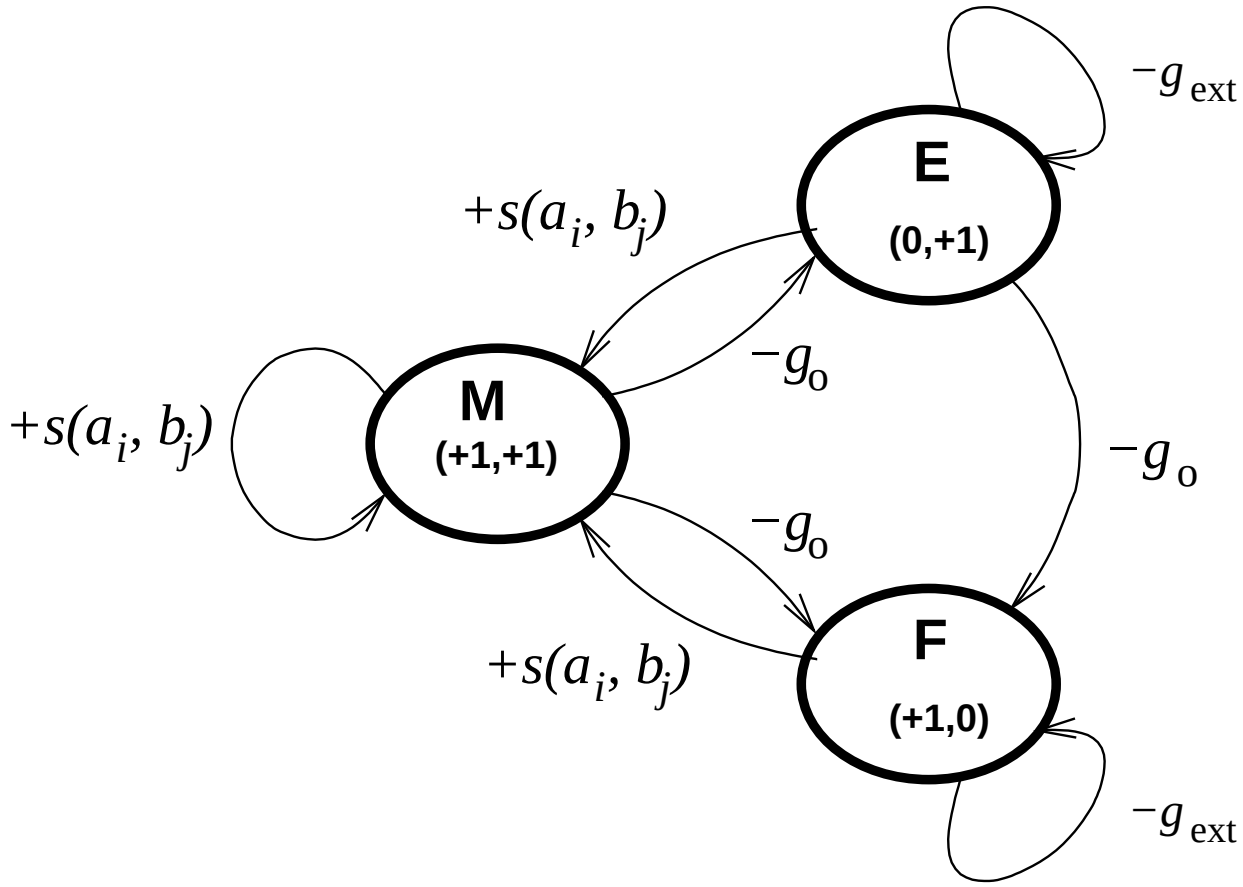
- linear: $\lambda(l_g) = -l_g d$

  l_g: length of the gap; d:gap penalty;

- affine: $\lambda(l_g) = -(g_o + g_{ext}(l_g - 1))$

  g_o: gap open and g_{ext}: gap extension penalty;

$$
\begin{aligned}
M(i,j) &= \max \begin{cases} M(i-1,j-1) + s(a_i, b_j) \\ E(i-1,j-1) + s(a_i, b_j) \\ F(i-1,j-1) + s(a_i, b_j) \end{cases} \\
E(i,j) &= \max \begin{cases} M(i,j-1) - g_o \\ E(i,j-1) - g_{ext} \end{cases} \\
F(i,j) &= \max \begin{cases} M(i-1,j) - g_o \\ E(i-1,j) - g_o \\ F(i-1,j) - g_{ext} \end{cases}
\end{aligned}
$$


E
(0,+1)
M
(+1,+1)
F
(+1,0)
$-g_{\text{ext}}$
$+s(a_i, b_j)$
$-g_{\text{o}}$
$+s(a_i, b_j)$
$-g_{\text{o}}$
$-g_{\text{o}}$
$+s(a_i, b_j)$
$-g_{\text{ext}}$

Alignment Scores: Substitutions

- Random Model: residues occure independently with some frequency q_a:

$$P(a, b \mid R) = q_a q_b$$

- Match Model: Aligned residue pairs occure with a joint probability p_{ab}:

$$P(a, b \mid M) = p_{ab}$$

- ratio of these two likelihoods: *odds ratio*:

$$\frac{P(a, b \mid M)}{P(a, b \mid R)} = \frac{p_{ab}}{q_a q_b} = p(a, b)$$

- additive scoring system: *log-odds ratio*

$$s(a, b) = \log\left(\frac{p_{ab}}{q_a q_b}\right)$$

  affine transformation of log likelihood ratio

$$s(a, b) = \omega + k \log\left(\frac{p_{ab}}{q_a q_b}\right)$$

Alignment Scores: Gaps (Indels)

- probability of a gap

$$p(\mathbf{g}) = f(l_g) \prod_{i \text{ in } \mathbf{g}} q_{a_i}$$

- odds ratio

$$\frac{p(\mathbf{g})}{\prod_{i \text{ in } \mathbf{g}} q_{a_i}} = \frac{f(l_g) \prod_{i \text{ in } \mathbf{g}} q_{a_i}}{\prod_{i \text{ in } \mathbf{g}} q_{a_i}} = f(l_g)$$

- additive scoring system: *log-odds ratio*

$$\gamma(l_g) = \log f(l_g)$$

  affine transformation

$$\gamma(l_g) = \omega + k \log f(l_g)$$

Probability of an alignment

Probability of a match

$$p_{ab} = p(a,b) q_a q_b$$

Probability of a gap

$$p(\mathbf{g}) = p_{-b_j} = p_{a_i -} = f(l_g) \prod_{i \text{ in } \mathbf{g}} q_{a_i}$$

Probability of the whole alignment

$$\begin{aligned} \text{Prob}(\mathcal{A}) &= \prod_{i \in \text{sub}} p_{a_i^* b_i^*} \prod_{i \in \text{ins}} p_{-b_i^*} \prod_{i \in \text{del}} p_{a_i^* -} = \\ &= p(\mathbf{a})\, p(\mathbf{b}) \prod_{i \in \text{indel}} f(l_{indel}) \prod_{i \in \text{sub}} p(a_i^*, b_i^*) \end{aligned}$$

$$
\begin{aligned}
\frac{\mathsf{Prob}(\mathcal{A})}{p(\mathbf{a})\,p(\mathbf{b})} &= \prod_{i\in\mathsf{indel}} f(l_{indel}) \prod_{i\in\mathsf{sub}} p(a_i^*, b_i^*) \\
S(\mathcal{A}) &= \omega + k\left\{ \sum_{l_g} \log f(l_g) + \sum_{(i,j)\in\mathcal{A}} \log p(a_i, b_j) \right\} \\
e^{S(\mathcal{A})} &= e^{\omega} \prod_{l_g} e^{k \log f(l_g)} \prod_{(i,j)\in\mathcal{A}} e^{k \log p(a_i,b_j)} \\
&= e^{\omega}\left\{ \prod_{l_g} f(l_g) \prod_{(i,j)\in\mathcal{A}} p(a_i, b_j) \right\}^k \\
&= e^{\omega}\left\{ \frac{\mathsf{Prob}(\mathcal{A})}{p(\mathbf{a})p(\mathbf{b})} \right\}^k \\
e^{S(\mathcal{A})/k} &= \frac{e^{(\omega/k)}}{p(\mathbf{a})p(\mathbf{b})} \mathsf{Prob}(\mathcal{A})
\end{aligned}
$$

Partition Function Z

$$
\begin{aligned}
c &= \frac{\mathrm{p}(\mathbf{a})\ \mathrm{p}(\mathbf{b})}{e^{(\omega/k)}} \\
\mathrm{Prob}(\mathcal{A}) &= c\, e^{\frac{S(\mathcal{A})}{k}} \\
\sum_{\mathcal{A}} \mathrm{Prob}(\mathcal{A}) &= 1 = c \sum_{\mathcal{A}} e^{\frac{S(\mathcal{A})}{k}}
\end{aligned}
$$

partition function Z: sum of the Boltzmann factors for all possible states

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

$c = \frac{1}{Z}$: relationship between "energies" and probabilities of states

$$\mathrm{Prob}(\mathcal{A}) = \frac{1}{Z} e^{\frac{S(\mathcal{A})}{k}}$$

Partition Function Z

- e.g. Dayhoffs PAM Matrices

$$\begin{aligned} s(a,b) &= 10 \log_{10} \left(\frac{p_{ab}}{q_a q_b} \right) \\ k &= \frac{10}{\log_e 10} \approx 4.3429 \end{aligned}$$

- temperature dependent partition function

$$Z(T) = \sum_{\mathcal{A}} e^{\frac{S(\mathcal{A})}{kT}} = \sum_{\mathcal{A}} e^{\beta S(\mathcal{A})}$$

Match Probabilities

$$
\begin{aligned}
\Omega_{i,j} &= \{\mathcal{A} \mid (i,j) \in \mathcal{A}\} \\
Prob(\Omega_{i,j}) &= \frac{1}{Z} \sum_{\mathcal{A} \in \Omega_{i,j}} e^{S_{\mathcal{A}}} = \frac{Z(\Omega_{i,j})}{Z} \\
S(\mathcal{A} \in \Omega_{i,j}) &= S(\mathcal{A}^{i,j}_{1,1}) + S(\mathcal{A}^{m,n}_{i,j}) - s(a_i, b_j)
\end{aligned}
$$

$$
\begin{aligned}
Z(\Omega_{i,j}) &= \sum_{\mathcal{A} \in \Omega_{i,j}} e^{\beta\, S(\mathcal{A}^{i,j}_{1,1}) + \beta\, S(\mathcal{A}^{m,n}_{i,j}) - \ beta\, s(a_i, b_j)} \\
&= \underbrace{\sum_{\mathcal{A} \in \mathfrak{A}^{i,j}_{1,1}} e^{\beta\, S(\mathcal{A}^{i,j}_{1,1})}}_{Z^{\mathsf{M}}_{ij}} \times \underbrace{\sum_{\mathcal{A} \in \mathfrak{A}^{m,n}_{i,j}} e^{\beta\, S(\mathcal{A}^{m,n}_{i,j})}}_{\widehat{Z}^{\mathsf{M}}_{ij}} \times\ e^{-\beta\, s(a_i, b_j)} \\
&= Z^{\mathsf{M}}_{ij}\ \widehat{Z}^{\mathsf{M}}_{ij}\ e^{-\beta\, s(a_i, b_j)}
\end{aligned}
$$

$$
prob_{(i,j)} = \frac{Z_{i,j} \widehat{Z}_{i,j}}{e^{s(a_i, b_j)} Z}
$$

Algorithm for the calculation of the partition function

$$
\begin{aligned}
Z^M_{i,j} &= \left(Z^M_{i-1,j-1} + Z^E_{i-1,j-1} + Z^F_{i-1,j-1} \right) e^{\beta s(a_i,b_i)} \\
Z^E_{i,j} &= Z^M_{i,j-1} e^{\beta g_o} + Z^E_{i,j-1} e^{\beta g_{\mathsf{ext}}} \\
Z^F_{i,j} &= \left(Z^M_{i-1,j} + Z^E_{i-1,j} \right) e^{\beta g_o} + Z^F_{i-1,j} e^{\beta g_{\mathsf{ext}}}
\end{aligned}
$$

$$
Z_{i,j} = Z^M_{i,j} + Z^E_{i,j} + Z^F_{i,j}
$$

Stochastic Backtracking

Probability of each state depends on the previous state

match

$$
\begin{aligned}
p(a,b) &= \frac{Z^M_{i-1,j-1}\, e^{\beta s(a_i,b_j)}}{Z^M_{i,j}} \\
p(-,b) &= \frac{Z^E_{i-1,j-1}\, e^{\beta s(a_i,b_j)}}{Z^M_{i,j}} \\
p(a,-) &= \frac{Z^F_{i-1,j-1}\, e^{\beta s(a_i,b_j)}}{Z^M_{i,j}}
\end{aligned}
$$

$$
\text{where } Z^M_{i,j} = (Z^M_{i-1,j-1} + Z^E_{i-1,j-1} + Z^F_{i-1,j-1}) e^{\beta s(a_i,b_j)}
$$

gap in **a**

$$
\begin{aligned}
p(a,b) &= \frac{Z^M_{i-1,j}\, e^{\beta g_o}}{Z^F_{i,j}} \\
p(-,b) &= \frac{Z^E_{i-1,j}\, e^{\beta g_o}}{Z^F_{i,j}} \\
p(a,-) &= \frac{Z^F_{i,j-1}\, e^{\beta g_{\text{ext}}}}{Z^F_{i,j}}
\end{aligned}
$$

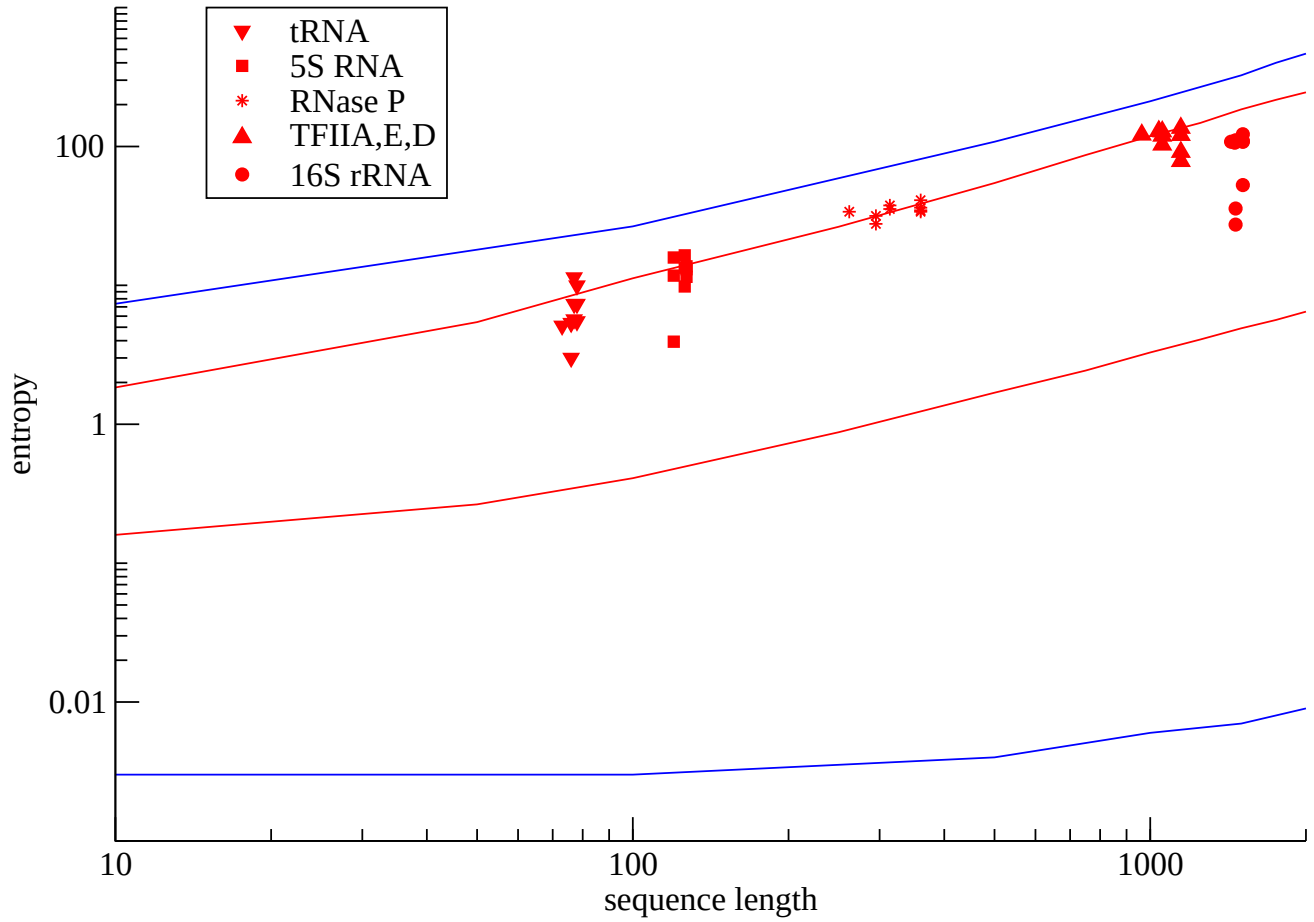
$$
\text{where } Z^F_{i,j} = (Z^M_{i-1,j} + Z^E_{i-1,j}) e^{\beta g_o} + Z^F_{i-1,j} e^{\beta g_{\text{ext}}}
$$

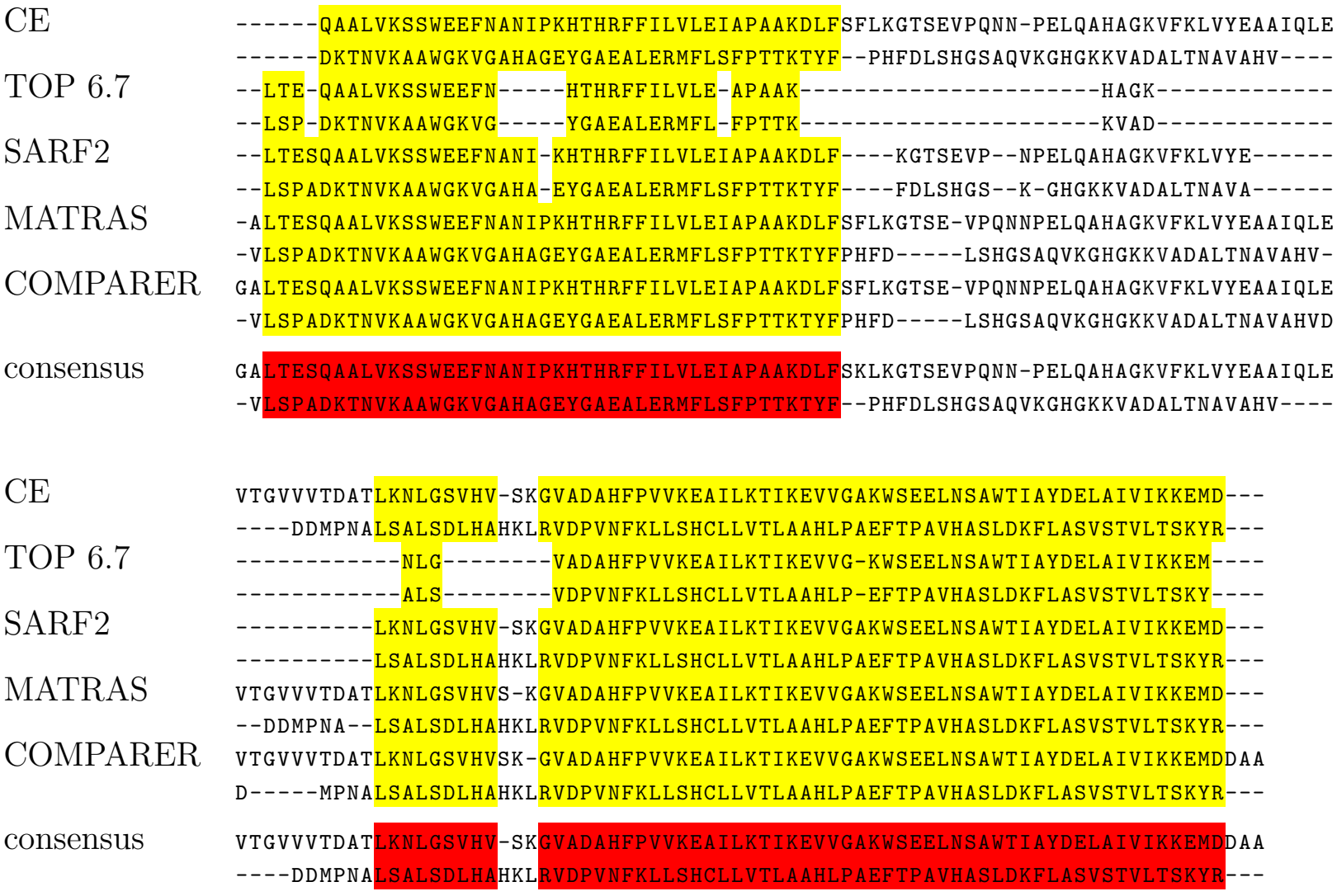
gap in **b**

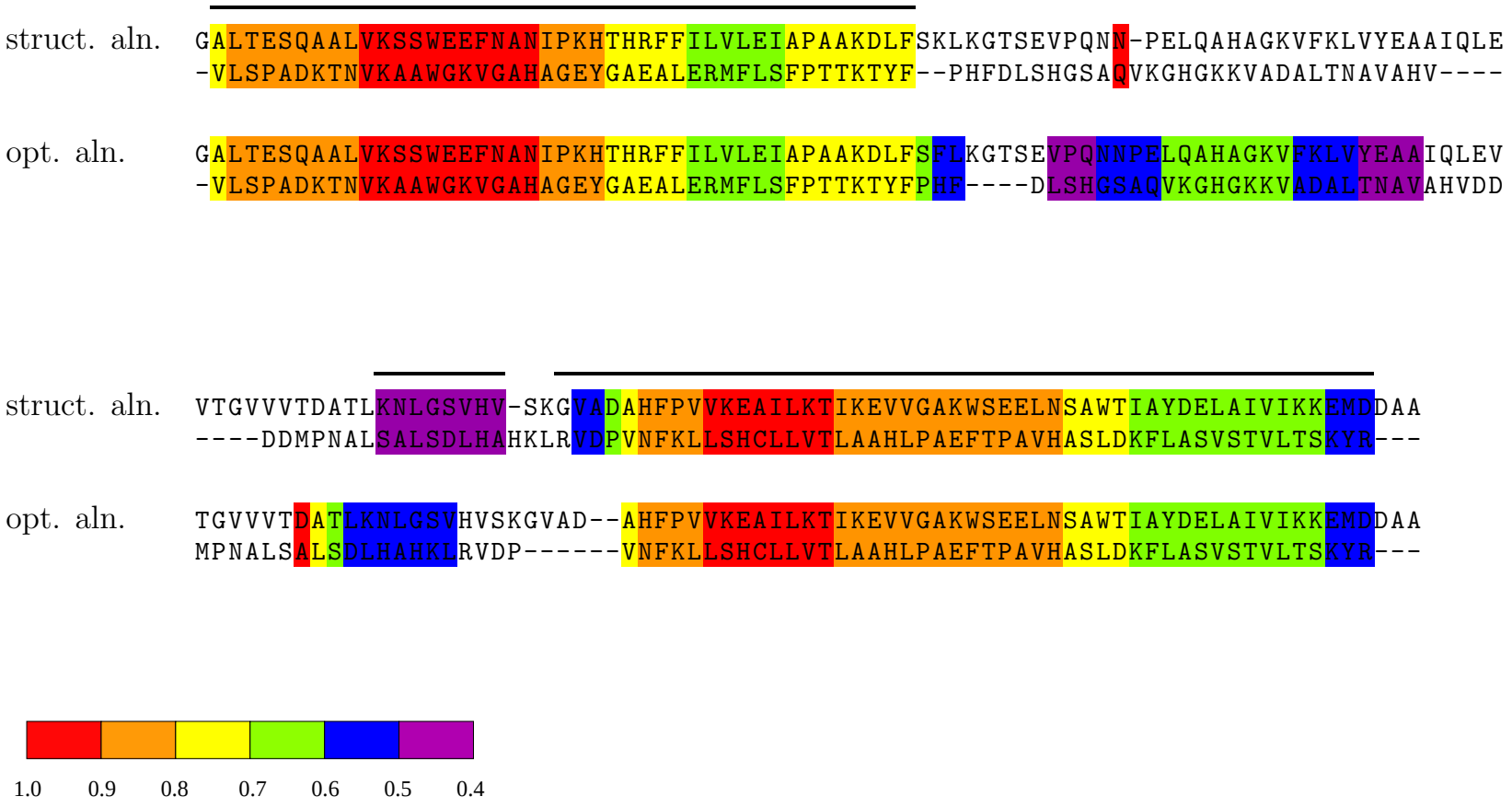
$$
\begin{aligned}
p(a,b) &= \frac{Z^M_{i,j-1}\, e^{\beta g_o}}{Z^E_{i,j}} \\
p(-,b) &= \frac{Z^E_{i,j-1}\, e^{\beta g_{\text{ext}}}}{Z^E_{i,j}}
\end{aligned}
$$

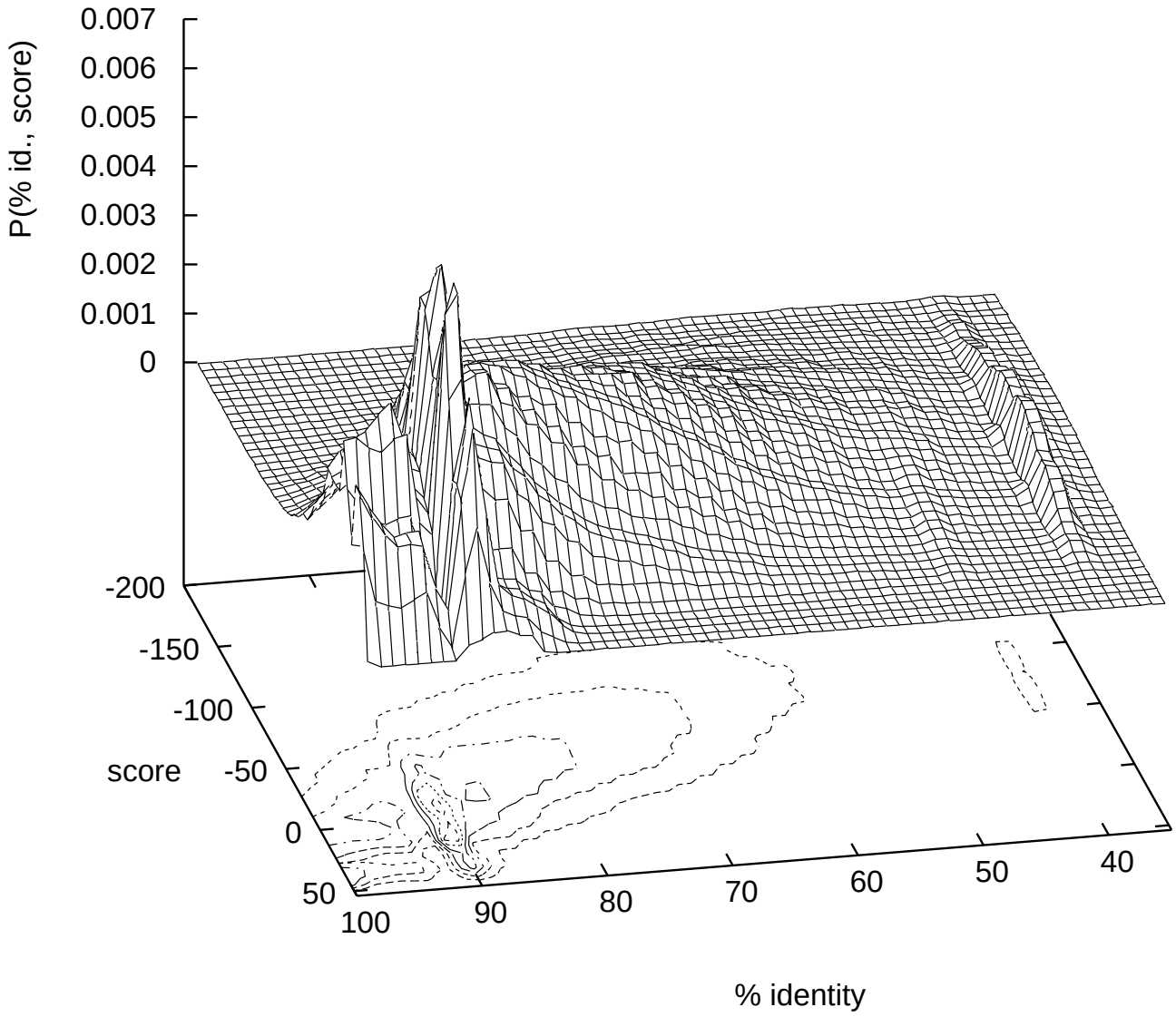
$$
\text{where } Z^E_{i,j} = Z^M_{i,j-1} e^{\beta g_o} + Z^E_{i,j-1} e^{\beta g_{\text{ext}}}
$$

$$\Delta S^{\mathsf{ensemble}} = S(\mathcal{A}_{\mathsf{opt}}) - kT\ln Z$$




CE
TOP 6.7
SARF2
MATRAS
COMPARER
consensus
CE
TOP 6.7
SARF2
MATRAS
COMPARER
consensus


struct. aln.
GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSKLKGTSEVPQNN-PELQAHAGKVFKLVYEAAIQLE
-VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYF--PHFDLSHGSAQVKGHGKKVADALTNAVAHV----
opt. aln.
GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSEVPQNNPELQAHAGKVFKLVYEAAIQLEV
-VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF----DLSHGSAQVKGHGKKVADALTNAVAHVDD
struct. aln.
VTGVVVTDATLKNLGSVHV-SKGVADAHFPVVKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMDDAA
----DDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR---
opt. aln.
TGVVVTDATLKNLGSVHVSKGVAD--AHFPVVKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMDDAA
MPNALSALSDLHAHKLRVDP------VNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR---
1.0
0.9
0.8
0.7
0.6
0.5
0.4


P(% id., score)
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0
-200
-150
-100
score
-50
0
50
100
90
80
70
60
50
40
% identity