ppALIGN: A tool to assess the uncertainty in score-based alignements

S. WOLFSHEIMER & G. NUEL

MAP5, Department of Applied Mathematics
University Paris Descartes,
Paris, France

January 18, 2010

Outline

1. Score-based alignment
   - Notion of alignment
   - Finding the best alignment
   - Example
2. Probabilistic alignment
   - Alignment biases
   - Gibbs-Boltzman distribution
3. Posterior probabilities
   - Examples
   - Partition function calculation
   - ppALIGN in action

Outline

1. Score-based alignment
   - Notion of alignment
   - Finding the best alignment
   - Example
2. Probabilistic alignment
   - Alignment biases
   - Gibbs-Boltzman distribution
3. Posterior probabilities
   - Examples
   - Partition function calculation
   - ppALIGN in action

How to compare two sequences ?

Problem

How to compare two biological sequences (DNA, proteins) $X = X_1 \ldots X_n$ and $Y = Y_1 \ldots Y_m$? Should we:

- compare their respective lengths ?
- compare their compositions (letters, word of size 2, 3, $\ldots$)
- look for repetitions ?
- $\ldots$

$\Rightarrow$ What about their proximity in the evolution process ?

Mutation and Indel

Definition

During the evolution of a biological sequence, a letter that changes is called a mutation and we called indel either the insertion of a letter or the deletion of a letter.

Example (a DNA sequence)

```
a  c  c  g  t  t  a     c  a  a  g  a  c  a
|  |  |  |  |  |  |     |  |  |  |  |  |  |
a  c  c  g  t  t  a     c  a  a  g  a  c  a
|  |  |  |  |  |  |     |  |  |  |  |  |  |
a  c  c  g  t  t  a     c  a  a  g  a  c  a
|  |  |  |  |  |  |     |  |  •  |  |  |  |
a  c  c  g  t  t  a     c  a  t  g  a  c  a
|  |  •  |  |  |  |     |  |  |  |  |  |  |
a  c     g  t  t  a     c  a  t  g  a  c  a
|  |     |  |  |  |  •  |  |  |  |  |  |  |
a  c     g  t  t  a  g  c  a  t  g  a  c  a
```

What is an alignment ?

Example (Two DNA sequences)

$$X = \mathrm{a\ c\ g\ t\ a\ g\ c\ a\ t\ g\ a\ c\ a}$$
$$Y = \mathrm{a\ c\ c\ g\ t\ a\ c\ a\ a\ g\ c\ a}$$

We denote by Z a common ancestral sequence:

Z	a	c	c	g	t	t	a		c	a	a	g	a	c	a
X	a	c		g	t		a	g	c	a	t	g	a	c	a
Y	a	c	c	g	t		a		c	a	a	g		c	a

here is the alignment we get:

$\widetilde{X}$	a	c	-	g	t	-	a	g	c	a	t	g	a	c	a
$\widetilde{Y}$	a	c	c	g	t	-	a	-	c	a	a	g	-	c	a

Outline

1. Score-based alignment
   - Notion of alignment
   - Finding the best alignment
   - Example
2. Probabilistic alignment
   - Alignment biases
   - Gibbs-Boltzman distribution
3. Posterior probabilities
   - Examples
   - Partition function calculation
   - ppALIGN in action

Scoring Alignments

Definition (Score of an Alignment)

Using the scoring function $\sigma : \mathcal{S} \cup \{\text{-}\} \times \mathcal{S} \cup \{\text{-}\} \to \mathbb{R}$ we define the score of an alignment as the sum of the scoring function over all the columns of the alignment.

Example ($\sigma(\text{match}) = +1$ $\sigma(\text{mismatch}) = -1$ $\sigma(\text{gap}) = -2$)

the first alignment scores $8 \times 1 - 0 \times 1 - 9 \times 2 = -10$

$\widetilde{X}$	a	c	-	g	t	a	-	-	-	g	c	a	t	g	a	c	a
$\widetilde{Y}$	a	c	c	g	t	a	c	a	a	g	c	a	-	-	-	-	-

and the second alignment scores $10 \times 1 - 1 \times 1 - 2 \times 2 = 5$

$\widetilde{X}$	a	c	-	g	t	a	g	c	a	t	g	a	c	a
$\widetilde{Y}$	a	c	c	g	t	a	-	c	a	a	g	-	c	a

How to find the best alignment ?

Brute force

Scores individually all possible alignments and pick up the best one.

$\Rightarrow$ How many possible alignments ?

Proposition

If $N_{n,m}$ is the number of alignments between two sequences of lengths n and m we have the following recurrence relation:

$$N(1,m) = 2m+1 \quad N(n,1) = 2n+1$$

$$N(n,m) = N(n,m-1) + N(n-1,m) + N(n-1,m-1)$$

for all $n, m \geqslant 1$.

Number of Alignments

Example

	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 6$	$m = 7$
$n = 1$	3	5	7	9	11	13	15
$n = 2$	5	13	25	41	61	85	113
$n = 3$	7	25	63	129	231	377	575
$n = 4$	9	41	129	321	681	1289	2241
$n = 5$	11	61	231	681	1683	3653	7183

Approximation

Idea: $N(n, m) \sim \rho^{n+m}$ we hence get

$$\rho^2 - 2\rho - 1 = 0 \Rightarrow \rho = 1 + \sqrt{2}$$

this gives us:

$$N(20, 20) \simeq 10^{15} \quad N(100, 100) \simeq 10^{76} \quad N(1000, 1000) \simeq 10^{764}$$

Dynamic Programming

Needleman and Wunsch (1970)

We denote by $B_{i,j}$ the best score of an alignment of $X_1 \dots X_i$ and $Y_1 \dots Y_j$ and we get

```
$B_{0,0} = 0$
$B_{i,0} = \sum_{k=1}^{i} \sigma(X_k, -)$ and $B_{0,j} = \sum_{k=1}^{j} \sigma(-, Y_k)$
**for** $i = 1 \dots n$ **do**
    **for** $i = 1 \dots m$ **do**
        $B_{i,j} = \max \begin{cases} B_{i-1,j-1} + \sigma(X_i, Y_j) \\ B_{i-1,j} + \sigma(X_i, -) \\ B_{i,j-1} + \sigma(-, Y_j) \end{cases}$
return $B_{n,m}$ and use a traceback to find the alignment
```

Outline

1. Score-based alignment
   - Notion of alignment
   - Finding the best alignment
   - Example
2. Probabilistic alignment
   - Alignment biases
   - Gibbs-Boltzman distribution
3. Posterior probabilities
   - Examples
   - Partition function calculation
   - ppALIGN in action

Example

Example (X = gcgacgtgcaag Y = aggcacgca $+3, -1, -2$)

	-	a	g	g	c	a	c	g	c	a
-	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
g	-2	-1	1	-1	-3	-5	-7	-9	-11	-13
c	-4	-3	-1	0	2	0	-2	-4	-6	-8
g	-6	-5	0	2	0	1	-1	1	-1	-3
a	-8	-3	-2	0	1	3	1	-1	0	2
c	-10	-5	-4	-2	3	1	6	4	2	0
g	-12	-7	-2	-1	1	2	4	9	7	5
t	-14	-9	-4	-3	-1	0	2	7	8	6
g	-16	-11	-6	-1	-3	-2	0	5	6	7
c	-18	-13	-8	-3	2	0	1	3	8	6
a	-20	-15	-10	-5	0	5	3	1	6	11
a	-22	-17	-12	-7	-2	3	4	2	4	9
g	-24	-19	-14	-9	-4	1	2	7	5	7

a	g	g	c	-	a	c	g	-	-	c	a	-	-
-	g	-	c	g	a	c	g	t	g	c	a	a	g

Outline

1. Score-based alignment
   - Notion of alignment
   - Finding the best alignment
   - Example
2. Probabilistic alignment
   - Alignment biases
   - Gibbs-Boltzman distribution
3. Posterior probabilities
   - Examples
   - Partition function calculation
   - ppALIGN in action

Alignment biases

- 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

Example (protein alignment with blosum62)

SALLASGGTSSHRWSRT	score = 31
SALLMARKSHRVLWSRT	
SALLASGGTSSHR--WSRT	score = 31
SALLMA--RKSHRVLWSRT	
SALLASGGTSSHR--WSRT	score = 28
SALL--MARKSHRVLWSRT	

Outline

1. Score-based alignment
   - Notion of alignment
   - Finding the best alignment
   - Example
2. Probabilistic alignment
   - Alignment biases
   - Gibbs-Boltzman distribution
3. Posterior probabilities
   - Examples
   - Partition function calculation
   - ppALIGN in action

Gibbs-Boltzman distribution

- Look at weighted space of all alignments
- Gibbs-Boltzmann distribution

$$P_T(\mathcal{A}) = \frac{1}{Z_T} \exp\left[\frac{1}{T} S(\mathcal{A})\right]$$

  temperature: T
  partition function: $Z_T = \sum_{\mathcal{A}} \exp\left[\frac{1}{T} S(\mathcal{A})\right]$

- $\lim_{T \to 0}$: $P_T(\mathcal{A}) = 1$ if $\mathcal{A}$ is optimal and 0 otherwise
- $\lim_{T \to \infty}$: $P_T(\mathcal{A}) = \text{const} \quad \forall \mathcal{A}$

[S. Miyazawa, Protein Eng. (1995)]

[M. Kschischo and M. Lässig, Pacific Symposium on Biocomputing 5 (2000)]

Probabilistic alignment

Example (weighted set of alignments T=1)

$\mathcal{A}$	$P(\mathcal{A})$	$\mathcal{A}$	$P(\mathcal{A})$

`A--CGT`
`|  |||   S = 1`
`ACGCGT`
	32.1 %	`A-CG-T`
		`| ++ |   S = -4`
		`ACGCGT`
			0.2 %

`A-C-GT`
`| + ||   S = -2`
`ACGCGT`
	1.6 %	`AC-G-T`
		`|| + |   S = -2`
		`ACGCGT`
			1.6 %

`AC--GT`
`||  ||   S = 1`
`ACGCGT`
	32.1 %	`ACG--T`
		`|||  |   S = 1`
		`ACGCGT`
			32.1 %

`A---CGT`
`|   + |   S = -5`
`ACGCG-T`
	0.8 %	`ACG---T`
		`| +   |   S = -5`
		`A-CGCGT`
			0.8 %

$$P(\mathcal{A}) = e^{S/T}/Z_T$$

$$Z_T = e^1 + e^{-2} + e^1 + e^{-5} + e^{-4} + e^{-2} + e^1 + e^{-5}$$

Probabilistic alignment

Example (weighted set of alignments T=10)

$\mathcal{A}$	$P(\mathcal{A})$	$\mathcal{A}$	$P(\mathcal{A})$
A--CGT | ||| S = 1 ACGCGT	16.2 %	A-CG-T | ++ | S = -4 ACGCGT	9.8 %
A-C-GT | + || S = -2 ACGCGT	12.0 %	AC-G-T || + | S = -2 ACGCGT	12.0 %
AC--GT || || S = 1 ACGCGT	16.2 %	ACG--T ||| | S = 1 ACGCGT	16.2 %
A---CGT | + | S = -5 ACGCG-T	8.9 %	ACG---T | + | S = -5 A-CGCGT	8.9 %

$$
\begin{aligned}
P(\mathcal{A}) &= e^{S/T}/Z_T \\
Z_T &= e^{10} + e^{-20} + e^{10} + e^{-50} + e^{-40} + e^{-20} + e^{10} + e^{-50}
\end{aligned}
$$

Outline

1. Score-based alignment
   - Notion of alignment
   - Finding the best alignment
   - Example

2. Probabilistic alignment
   - Alignment biases
   - Gibbs-Boltzman distribution

3. Posterior probabilities
   - Examples
   - Partition function calculation
   - ppALIGN in action

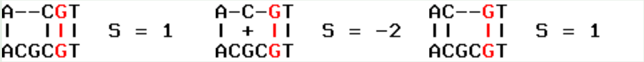
Posterior probabilities

- Alignment accuracy = posterior probabilities:

$$P\left[\begin{array}{ccc} \cdots & X_i & \cdots \\ \cdots & Y_j & \cdots \end{array}\right] = \frac{1}{Z_T} \sum_{\mathcal{A}:(X_i,Y_j)\in\mathcal{A}} \exp[S(\mathcal{A})/T]$$

Example (A pair, T=1)

$$P\left[\begin{array}{ccc} \cdots & G_3 & \cdots \\ \cdots & G_5 & \cdots \end{array}\right] = (e^1 + e^{-2} + e^1)/Z_T \approx 65.9\%$$



```
A--CGT          A-C-GT           AC--GT
|  |||  S = 1   | + ||  S = -2   ||  ||  S = 1
ACGCGT          ACGCGT           ACGCGT
```

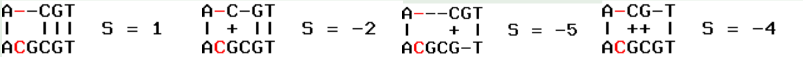
Posterior probabilities

- Alignment accuracy = posterior probabilities:

$$P\begin{bmatrix} \cdots & X_i & \cdots \\ \cdots & Y_j & \cdots \end{bmatrix} = \frac{1}{Z_T} \sum_{\mathcal{A}:(X_i,Y_j)\in\mathcal{A}} \exp[S(\mathcal{A})/T]$$

Example (A gap, T=1)

$$P\begin{bmatrix} \cdots & - & \cdots \\ \cdots & C_2 & \cdots \end{bmatrix} = (e^1 + e^{-2} + e^{-5} + e^{-4})/Z_T \approx 34.0\%$$



```
A--CGT              A-C-GT               A---CGT               A-CG-T
|  |||   S = 1      | + ||   S = -2      |   + |   S = -5      | ++ |   S = -4
ACGCGT              ACGCGT               ACGCG-T               ACGCGT
```

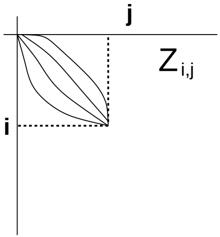
Outline

1. Score-based alignment
   - Notion of alignment
   - Finding the best alignment
   - Example
2. Probabilistic alignment
   - Alignment biases
   - Gibbs-Boltzman distribution
3. Posterior probabilities
   - Examples
   - Partition function calculation
   - ppALIGN in action

Partition function calculation

- $B_{i,j} \to Z_{i,j}$, max $\to \sum$ and $+ \to \times$
- Recall: $Z_T = \sum_{\mathcal{A}} e^{S(\mathcal{A})/T}$
- Forward algorithm
  $Z_{i,j}$: sum over all alignments of $X_1 \dots X_i$ and $Y_1 \dots Y_j$
- Here: $T = 1$

$$\begin{aligned} Z_{i,j} &= \underbrace{Z_{i-1,j-1}\, e^{\sigma(X_i,Y_j)}}_{\text{match/mismatch}} \\ &\quad + \underbrace{Z_{i-1,j}\, e^{\sigma(X_i,-)} + Z_{i,j-1}\, e^{\sigma(-,Y_j)}}_{\text{gap}} \\ Z_T &= Z_{n,m} \end{aligned}$$

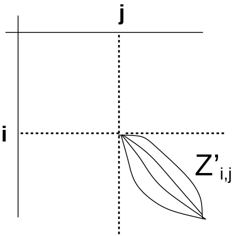


Partition function calculation

- $B_{i,j} \to Z_{i,j}$, $\max \to \sum$ and $+ \to \times$
- Recall: $Z_T = \sum_{\mathcal{A}} e^{S(\mathcal{A})/T}$
- Backward algorithm
  $Z'_{i,j}$: sum over all alignments of $X_{i+1} \dots X_n$ and $Y_{j+1} \dots Y_m$
- Here: $T = 1$

$$Z'_{i,j} = \underbrace{Z'_{i+1,j+1}\, e^{\sigma(X_{i+1},Y_{j+1})}}_{\text{match/mismatch}} + \underbrace{Z'_{i+1,j}\, e^{\sigma(X_{i+1},-)} + Z'_{i,j+1}\, e^{\sigma(-,Y_{j+1})}}_{\text{gap}}$$

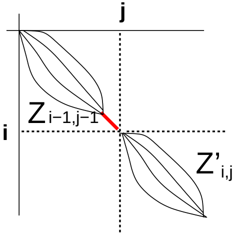
$$Z'_T = Z'_{0,0} \equiv Z_{n,m}$$



Partition function calculation

- $B_{i,j} \to Z_{i,j}$, max $\to \sum$ and $+ \to \times$
- Recall: $Z_T = \sum_{\mathcal{A}} e^{S(\mathcal{A})/T}$
- Combining forward and backward
- Here: $T = 1$

$$P\begin{bmatrix} \dots & X_i & \dots \\ \dots & Y_j & \dots \end{bmatrix} = \frac{Z_{i-1,j-1}\, e^{\sigma(X_i,Y_j)}\, Z'_{i,j}}{Z_T}$$



Outline

1. Score-based alignment
   - Notion of alignment
   - Finding the best alignment
   - Example

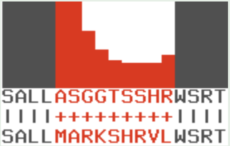
2. Probabilistic alignment
   - Alignment biases
   - Gibbs-Boltzman distribution

3. Posterior probabilities
   - Examples
   - Partition function calculation
   - ppALIGN in action

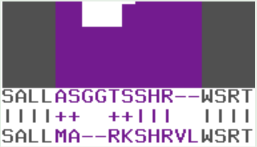
Posterior probabilities

Example (protein alignment revisited)

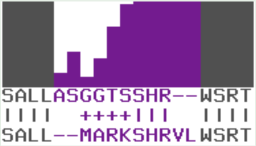
Alignment	Score
SALLASGGTSSHRWSRT SALLMARKSHRVLWSRT	score = 31
SALLASGGTSSHR--WSRT SALLMA--RKSHRVLWSRT	score = 31
SALLASGGTSSHR--WSRT SALL--MARKSHRVLWSRT	score = 28



score = 31



score = 31



score = 28

Our software

Generalizations / other features

- Affine gap costs: $g(l) = \alpha + \beta l$ instead of $g(l) = \gamma l$
- Pair Hidden Markov models.
  Probabilistic description without temperature parameter.
- Other decoding algorithms than optimal alignments, Sampling from Gibbs-Boltzmann distribution
  $P_T[\mathcal{A}] \propto e^{S(\mathcal{A})/T}$
- Different alignment models (e.g. global or local)

The ppALIGN webserver

- Standalone programs and C++ library (open source)
- Webinterface

http://www.math-info.univ-paris5.fr/ppblast/

Summary / Outlook

Summary

- Score based alignment common tool
- Alignment biases: many high scoring alignments
- Probabilistic description of alignment space. Posterior probabilities, alternative alignments.
- ppALIGN: software to analyse alignments

Outlook

- Parameter optimization (choice of temperature).
- Distribution of pattern in alignments
- Biological applications where correct alignment close to gaps is crucial