PALM: Probabilistic Area Loss Minimization for Protein Sequence Alignment

Fan Ding[a], Nan Jiang[a], Jianzhu Ma, Jian Peng, Jinbo Xu and Yexiang Xue
{ding274, jiang631, yexiang}@purdue.edu, majianzhu@pku.edu.cn, jianpeng@illinois.edu, jinboxu@gmail.com

[a] These authors contribute equally.

The Importance of this Problem

- End-to-End style learning for aligning proteins without the repeated workload.
- The algorithm can do robust learning to reduce the noises in the Biological dataset.
- The developed algorithm can help to find new proteins and drug discovery.

Pairwise Protein Alignment Problem

Given a sequence pair (S, T), $S = SLA$, $T = LRP$ and $a = [I_S, I_T, M, I_T, I_S]$.

1. Symbols M, I_S and I_T: represent a match, an insertion in S, and an insertion in T, respectively.
2. Alignment a: a sequential symbols M, I_S and I_T.

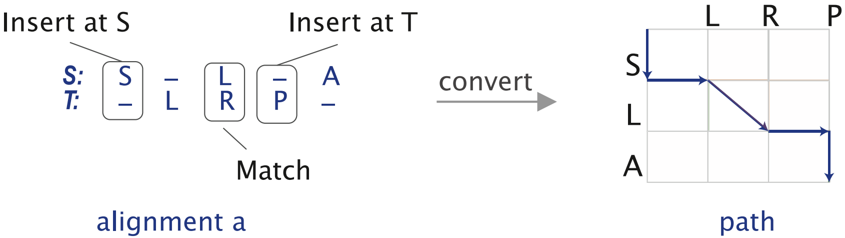
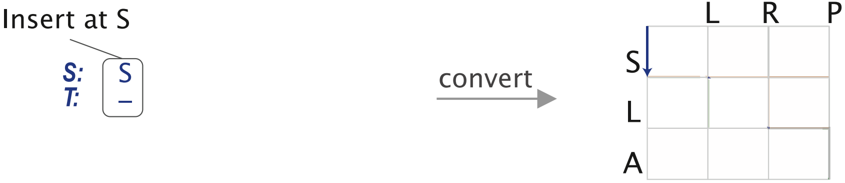


Figure: Alignment Matrix of sequence pair (S, T).

Alignment Sequence as Path

Given a sequence pair (S, T), $S = SLA$ and $T = LRP$.

$a = [I_S,]$



alignment a

path

Figure: Alignment Matrix of sequence pair (S, T).

Alignment Sequence as Path

Given a sequence pair (S, T), $S = SLA$ and $T = LRP$.

$a = [I_S, I_T]$

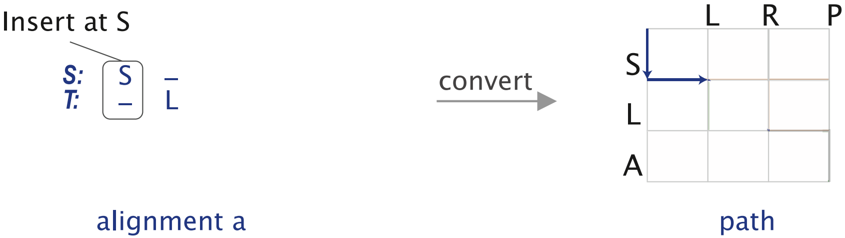


Figure: Alignment Matrix of sequence pair (S, T).

Alignment Sequence as Path

Given a sequence pair (S, T), $S = SLA$ and $T = LRP$.

$a = [I_S, I_T, M]$

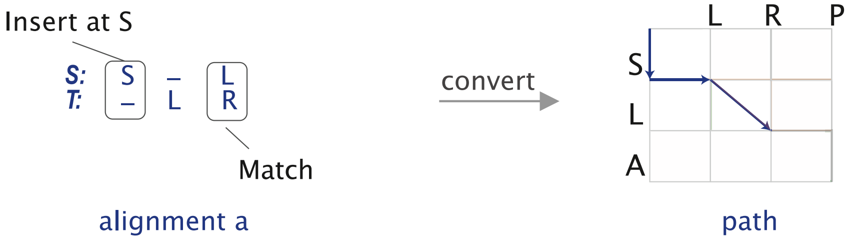


Figure: Alignment Matrix of sequence pair (S, T).

Alignment Sequence as Path

Given a sequence pair (S, T), $S = SLA$ and $T = LRP$.

$a = [I_S, I_T, M, I_T]$

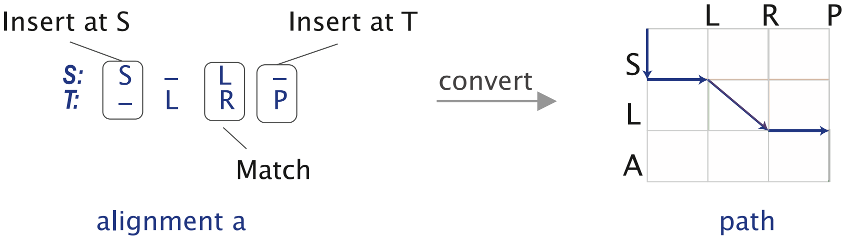


Figure: Alignment Matrix of sequence pair (S, T).

Alignment Sequence as Path

Given a sequence pair (S, T), $S = SLA$ and $T = LRP$.

$a = [I_S, I_T, M, I_T, I_S]$

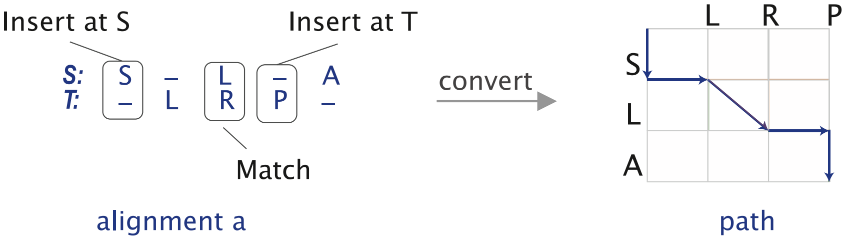


Figure: Alignment Matrix of sequence pair (S, T).

Pairwise Protein Alignment Problem

Given a sequence pair (S, T), $S = SLA$, $T = LRP$ and $a = [I_S, I_T, M, I_T, I_S]$.

We need $Pr_\theta(a|S, T)$: the probability of alignment a with parameter θ.

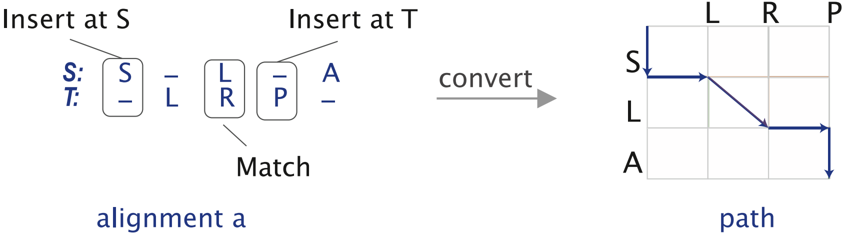


Figure: Alignment Matrix of sequence pair (S, T).

Our main tasks

- Learning. Given a training set $\{S, T, a^*\}$, we learn

$$\max_{\theta} Pr_{\theta}(a^*|S, T)$$

- Inference. Given two new sequence S', T', predict the most likely alignment $\hat{a}$:

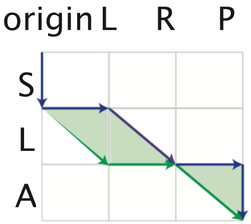
$$\hat{a} = \arg\max_{a \in \mathcal{A}} Pr(a|S', T')$$

Our Observations

- Biology datasets contain notable errors and alignment offsets from the real experiments.
- Existing approaches are not robust. Because they minimize of the pointwise differences of the two alignments.
- We consider a metric over the area of two alignments.

Example for falling of pointwise loss

gt	*S:*	S	–	L	–	A
	T:	–	L	R	P	–
$pred_1$	*S:*	S	L	–	A	
	T:	–	L	R	P	

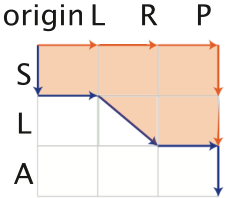


$$\mathcal{L}_{point}\left(gt, pred_1\right) = 4/5$$

Figure: Point-wise loss between ground-truth and pred_1.

Example for falling of pointwise loss

gt	S:	S	–	L	–	A	
	T:	–	L	R	P	–	
pred$_2$	S:	–	–	–	S	L	A
	T:	L	P	R	–	–	–

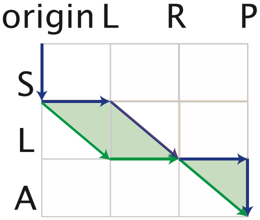


$$\mathcal{L}_{point}(gt, pred_2) = 4/5$$

Figure: Point-wise loss between ground-truth and pred$_2$.

Example for area loss

gt	***S:***	S	–	L	–	A
	T:	–	L	R	P	–
$pred_1$	***S:***	S	L	–	A	
	T:	–	L	R	P	

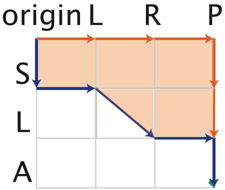


$$\mathcal{L}_{area}\left(gt, pred_1\right) = 3 \big/ 2$$

Figure: Area loss between ground-truth and pred_1.

Example for area loss

gt	*S:*	S	–	L	–	A	
	T:	–	L	R	P	–	
$pred_2$	*S:*	–	–	–	S	L	A
	T:	L	P	R	–	–	–



$$\mathcal{L}_{area}(gt, pred_2) = 4 + 1/2$$

Figure: Area loss between ground-truth and pred_2.

Probabilistic Area Distance via MRF I

Our original goal is to :

$$\max_{\theta} Pr_{\theta}\left(a^{*}|S,T\right)$$

With the integration of area loss, we extend to:

$$\max Pr(a^{*}|S,T) = \max \sum_{a} Pr_{area}(a^{*}|a,S,T)Pr_{\theta}(a|S,T). \tag{1}$$

which sums over the latent variable a.

Lower bound for efficient learning

- Learning efficiency concern: sums over latent alignments $a \in \mathscr{A}$ is exponential complex;

We use the lower bound

$$\hat{a} = \arg\max_{a \in \mathscr{A}} Pr_{area}(a^*|a, S, T) Pr_\theta(a|S, T) \tag{2}$$

$$Pr_{LB}(a^*|S, T) \approx Pr_{area}(a^*|\hat{a}, S, T) Pr_\theta(\hat{a}|S, T). \tag{3}$$

because of the principle of log-sum-exp function: summation usually dominated by one alignment.

Overview

Training:

1. get sample $\{S, T, a\}$.
2. compute $Pr_{LB}(a^*|S, T)$.
3. sample alignments for computing gradients (see details in paper).
4. repeat 1-3 training until converge.

Testing:

1. given S', T', predict $\hat{a}$ by:

$$\arg\max_{a \in \mathcal{A}} Pr(a|S', T')$$

Precision, Recall and F1-Score Benchmark

1. Sequence S length is between $[1, 100]$; Sequence T length is between $[100, 200]$;
2. "exact": only an exactly matched alignment is used for computing the true positive rate.

	$\lvert S\rvert \in [1, 100], \lvert T\rvert \in [100, 200]$		
	Precision (%)	Recall (%)	F1-Score (%)
	exact	exact	exact
DP	7.8	20.4	11.3
PALM	**9.9**	**23.5**	**13.9**

Table: PALM gets better results especially on longer sequences and remote homologies than the competing approach.

Precision, Recall and F1-Score Benchmark

1. "4-offset" scenario is a relaxed measure that 4-position off the exact match is allowed.
2. "10-offset" case is relaxed measure with 10-position off.

	$\lvert S\rvert \in [1, 100], \lvert T\rvert \in [100, 200]$								
	Precision (%)			Recall (%)			F1-Score (%)		
	exact	4off	10off	exact	4off	10off	exact	4off	10off
DP	7.8	**31.3**	**51.2**	20.4	39.0	56.3	11.3	34.7	53.6
PALM	**9.9**	29.8	48.7	**23.5**	**43.1**	**62.3**	**13.9**	**35.2**	**54.7**

Table: PALM gets better results on related measurements with "4-offset" and "10-offset".

Precision, Recall and F1-Score Benchmark

	$\lvert S\rvert \in [1, 100], \lvert T\rvert \in [100, 200]$									$\lvert S\rvert \in [100, 200], \lvert T\rvert \in [1, 100]$								
	Precision (%)			Recall (%)			F1-Score (%)			Precision (%)			Recall (%)			F1-Score (%)		
	exact	4-off	10-off	exact	4-off	10-off	exact	4-off	10-off	exact	4-off	10-off	exact	4-off	10-off	exact	4-off	10-off
DP	7.8	**31.3**	**51.2**	20.4	39.0	56.3	11.3	34.7	53.6	20.2	40.4	59.4	6.1	26.3	**45.1**	9.4	31.9	51.3
PALM	**9.9**	29.8	48.7	**23.5**	**43.1**	**62.3**	**13.9**	**35.2**	**54.7**	**26.8**	**44.6**	**63.2**	**6.4**	**26.6**	43.1	**10.3**	**33.3**	**51.2**

	$\lvert S\rvert \in [100, 200], \lvert T\rvert \in [400, +\infty)$									$\lvert S\rvert \in [400, +\infty), \lvert T\rvert \in [100, 200]$								
	Precision (%)			Recall (%)			F1-Score (%)			Precision (%)			Recall (%)			F1-Score (%)		
	exact	4-off	10-off	exact	4-off	10-off	exact	4-off	10-off	exact	4-off	10-off	exact	4-off	10-off	exact	4-off	10-off
DP	4.9	**24.1**	**41.0**	33.4	38.1	42.6	8.5	29.5	41.8	34.9	39.9	44.6	2.8	**14.4**	**24.8**	5.2	21.2	31.9
PALM	**6.1**	23.4	38.3	**61.1**	**69.0**	**76.5**	**11.1**	**34.9**	**51.0**	**62.5**	**71.0**	**78.8**	**3.2**	14.1	23.6	**6.1**	**23.5**	**36.3**

Table: PALM result for two testing sets with different lengths.

Conclusion

- We propose robust method for reducing the biological errors and offsets for Protein Alignment.
- We derive efficient dynamic sampling algorithm for model training.
- We demonstrate superior performance against competing approach over Precision/Recall/F1-score.

Q & A