A Convex Atomic-Norm Approach to Multiple Sequence Alignment & Motif Discovery

Ian E.H. Yen

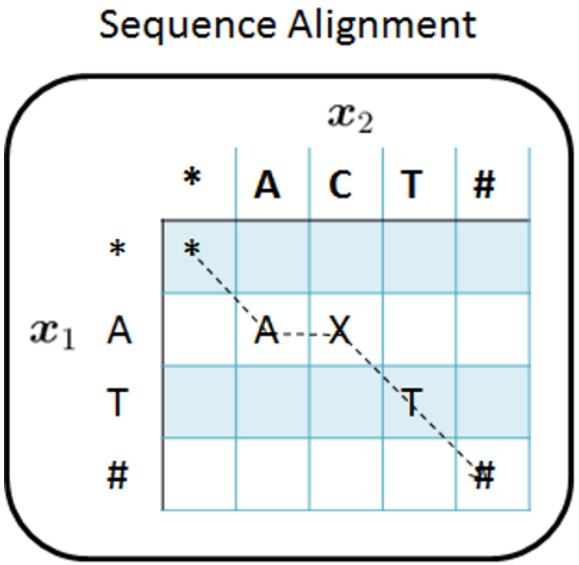
Joint work with Xin Lin, Jiong Zhang,
Pradeep Ravikumar and Inderjit Dhillon

Department of Computer Science
University of Texas at Austin

Outline

1. Multiple Sequence Alignment & Motif Discovery
2. Convex Relaxation via Atomic Constraints
3. Greedy Direction Method of Multiplier (GDMM)
4. Experimental Results

Multiple Sequence Alignment

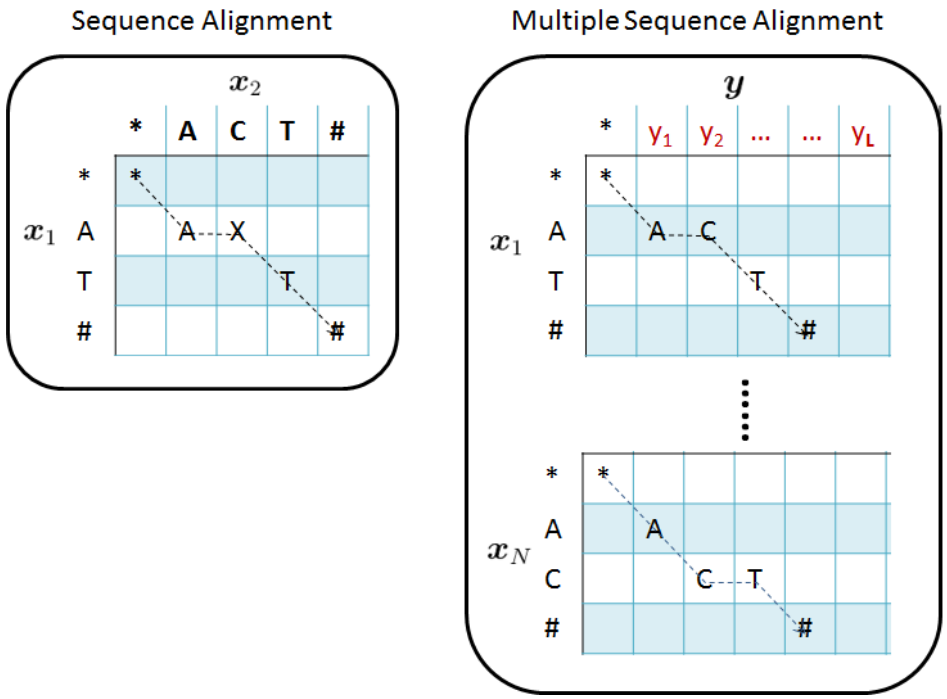


- **Sequence Alignment**: An alignment a is a path of transitions $t_1,..,t_K$ between states $(i,j) \in [\ell_1] \in [\ell_2]$ of reads on two sequences $\boldsymbol{x}_1$, $\boldsymbol{x}_2$.
  The sequence alignment problem can be expressed as

$$a^* = arg\min_a \; d(a; x_1, x_2) := \sum_{t \in a} d(t; x_1, x_2).$$

  where transition $t \in \{\text{matching, insertion, deletion}\}$.

Multiple Sequence Alignment

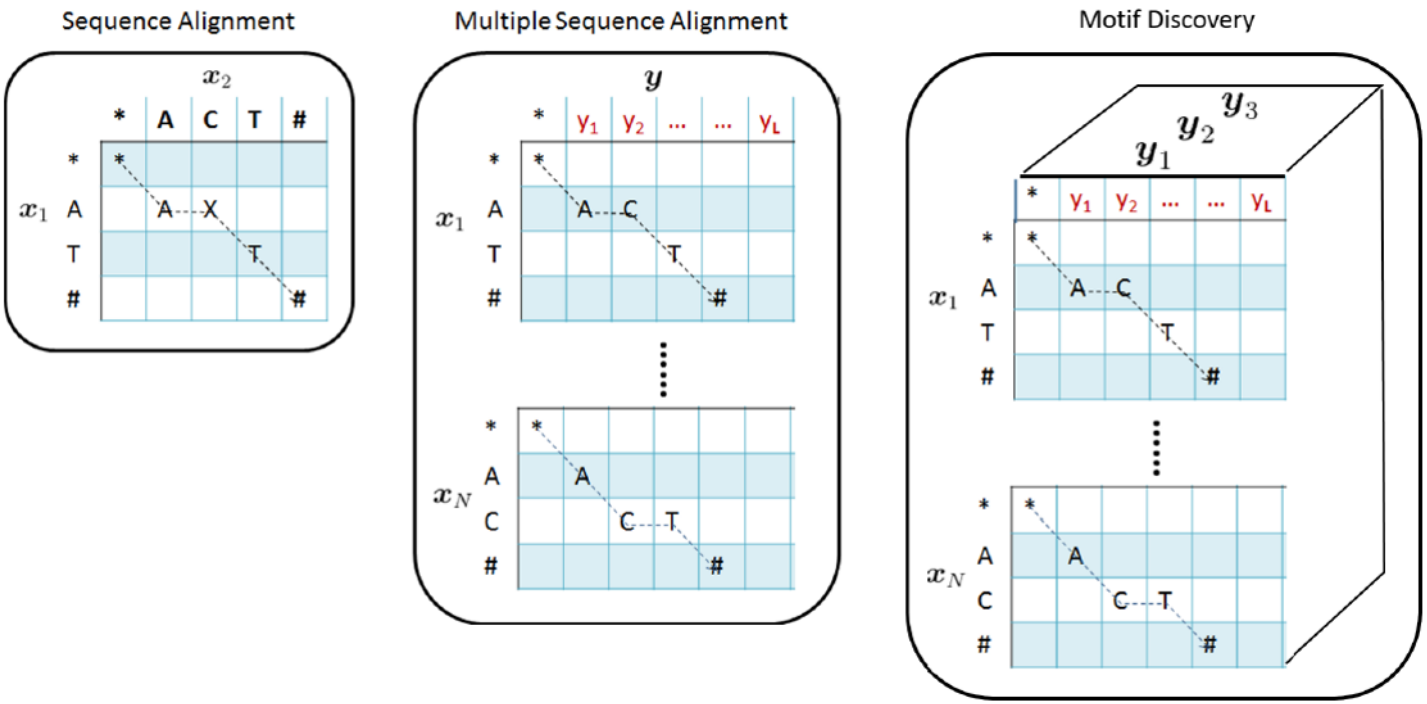


- **Multiple Sequence Alignment (MSA)**:

$$(y^*, a_1^*, ..., a_N^*) = \underset{y, a_1, ..., a_N}{argmin} \sum_{n=1}^{N} d(a_n; x_n, y). \tag{1}$$

aims to find a latent consensus sequence y^* and alignments $a_1, ..., a_N$ jointly.
(1) is called Star Alignment in distinction to Sum-of-Pairs objective, both of which are NP-hard.

Motif Discovery

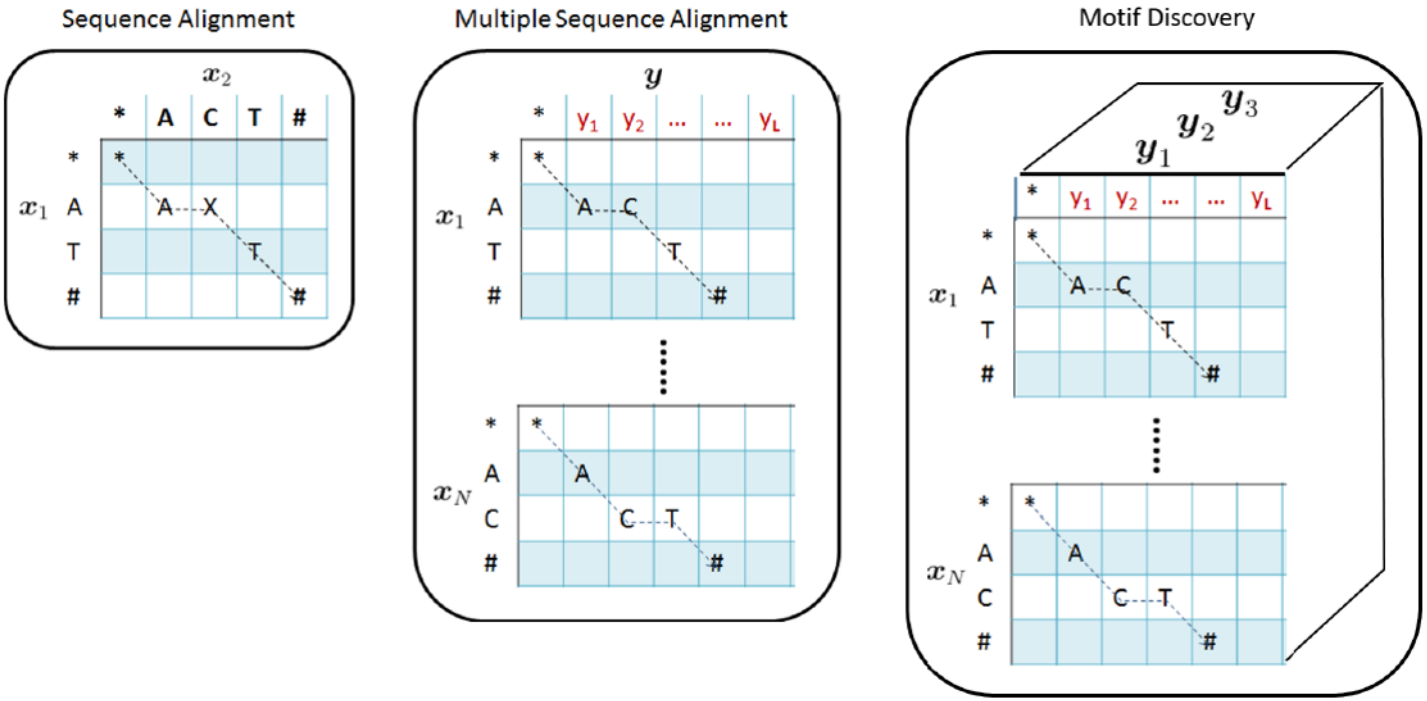


- **Motif Discovery (MD)**:

$$(y_1^*, ..., y_K^*; a_1^*, ..., a_N^*) := \underset{y,a}{argmin} \sum_{n=1}^{N} d(a_n; x_n, y_1, ..., y_K) \tag{2}$$

is a further generalization of MSA, where multiple motifs y_1,...,y_K can be aligned to segments of sequence. Typically, an insertion is only considered as gap between motifs (not inside motif).

Existing Approaches



- HMM: Model with (Profile) HMM and estimate via EM-style algorithm.
- Progressive Method: Reduce MSA to a series of Pairwise SA.
- Iterative, hill-climbing methods.

Outline

1. Multiple Sequence Alignment & Motif Discovery
2. Convex Relaxation via Atomic Constraints
3. Greedy Direction Method of Multiplier (GDMM)
4. Experimental Results

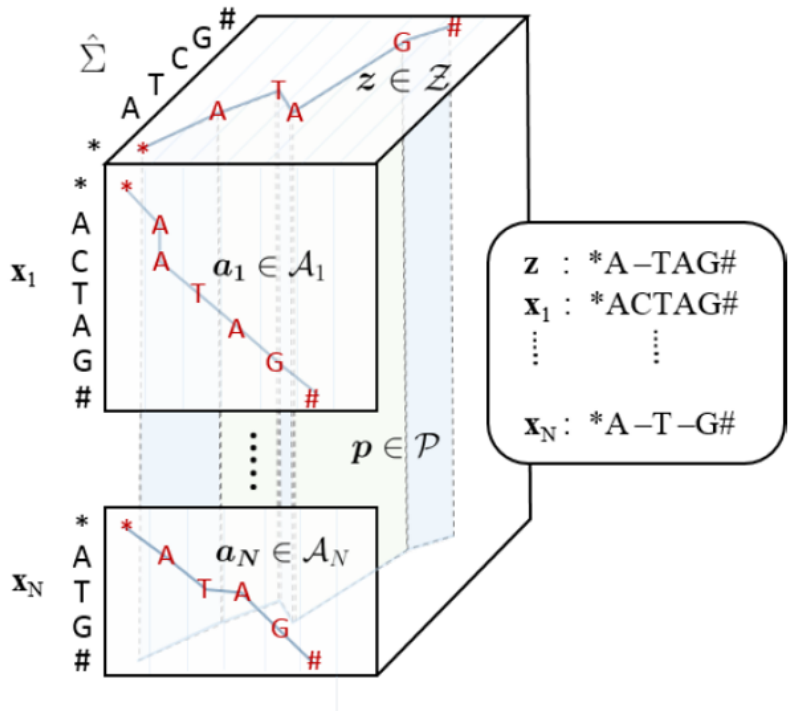
Convex Relaxation via Atomic Constraints: MSA

- MSA via Atomic Sets:

$$\begin{aligned} \min_{W \in \mathcal{M}} \quad & \langle D, W \rangle \\ s.t. \quad & W_n \in \mathcal{A}_n \\ & W \in \mathcal{P}. \end{aligned}$$

- Convex Relaxation:

$$\begin{aligned} \min_{W \in \mathcal{M}_{\mathbb{R}}} \quad & \langle D, W \rangle \\ s.t. \quad & W_n \in Conv(\mathcal{A}_n) \\ & W \in Conv(\mathcal{P}). \end{aligned}$$



- $\mathcal{M} = \ell \times L \times |\hat{\Sigma}|$.
- $\mathcal{A}_n$: all possible alignments.
- $\mathcal{P}$: any "folding screen" $\boldsymbol{p}$ of single consensus $\boldsymbol{z}$.

Convex Relaxation via Atomic Constraints: MD

- MD uses representation similar to MSA, with 3rd dimension $|\hat{\Sigma}|$ replaced by the number of all possible motifs $\sum_{L=L_{\min}}^{L_{\max}} |\Sigma|^L$.

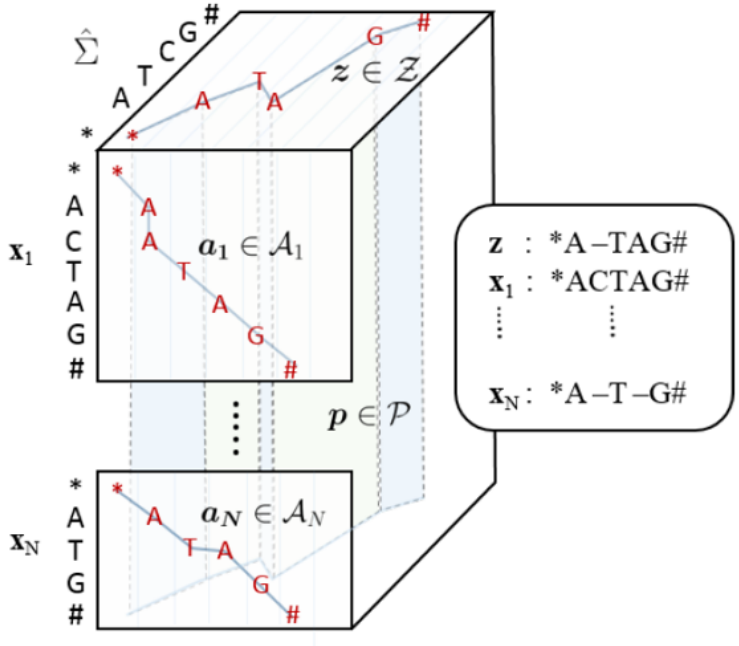
- Replace $W \in Conv(\mathcal{P})$ with the atomic-norm constraint $\mathbf{\Omega}_{\mathcal{P}_C}(\mathbf{W}) \leq \mathbf{K}$, where $\mathbf{\Omega}_{\mathcal{P}_C}(\mathbf{W}) := \mathbf{inf}\{\mathbf{q} \geq \mathbf{0} : \mathbf{W} \in \mathbf{q} * \mathbf{Conv}(\mathcal{P})\}$.

- MD via Atomic Sets:

$$\begin{aligned} \min_{W \in \mathcal{M}} \quad & \langle D, W \rangle \\ s.t. \quad & W_n \in \mathcal{A}_n \\ & \Omega_{\mathcal{P}_C}(W) = K. \end{aligned}$$

- Convex Relaxation:

$$\begin{aligned} \min_{W \in \mathcal{M}_{\mathbb{R}}} \quad & \langle D, W \rangle \\ s.t. \quad & W_n \in Conv(\mathcal{A}_n) \\ & \Omega_{\mathcal{P}_C}(W) \leq K. \end{aligned}$$



Outline

1. Multiple Sequence Alignment & Motif Discovery
2. Convex Relaxation via Atomic Constraints
3. Greedy Direction Method of Multiplier (GDMM)
4. Experimental Results

Greedy Direction Method of Multiplier (GDMM)

- To decouple the two atomic constraints of the convex relaxation, we minimize Augmented Lagrangian (AL) of

$$\begin{aligned} \min_{W_1, W_2 \in \mathcal{M}_{\mathbb{R}}} \quad & \langle D, W_1 \rangle + \frac{\rho}{2} \| W_1 - W_2 \|^2 \\ s.t. \quad & W_1 \in Conv(\mathcal{A}) \\ & W_2 \in Conv(\mathcal{P}) \\ & W_1 = W_2. \end{aligned}$$

w.r.t. W_1, W_2 separately, followed by a Dual Ascent step

$$Y^{(t+1)} = Y^{(t)} + \eta \left(W_1^{(t+1)} - W_2^{(t+1)} \right),$$

where Y is dual variable corresponding to constraint $W_1 = W_2$.

Greedy Direction Method of Multiplier (GDMM)

- Minimizing Augmented Lagrangian

$$\mathcal{L}(W_1, W_2, Y) := \langle D, W_1 \rangle + \langle Y, W_1 - W_2 \rangle + \frac{\rho}{2} \| W_1 - W_2 \|^2$$

  exactly w.r.t. W_1 or W_2 is hard, due to complex atomic constraint.

- We propose a GDMM algorithm, which minimizes (W_1, W_2) via only a (non-drop) step of Away-step Frank-Wolfe for each Dual Ascent step.

- The Frank-Wolfe step only requires computation of the greedy atoms:

$$W_1^{FW} := \underset{W_1 \in Conv(\mathcal{A})}{argmin} \langle D + \rho(W_1 - W_2) + Y, W_1 \rangle$$

$$W_2^{FW} := \underset{W_2 \in Conv(\mathcal{P})}{argmax} \langle \rho(W_1 - W_2) + Y, W_2 \rangle$$

  using local linear approximation, which can be solved via Smith-Waterman alignment and Viterbi Algorithm respectively.

- We show that GDMM converges to ϵ suboptimality in $O(1/\epsilon)$ iterations.

Outline

1. Multiple Sequence Alignment & Motif Discovery
2. Convex Relaxation via Atomic Constraints
3. Greedy Direction Method of Multiplier (GDMM)
4. Experimental Results

Experimental Result: Multiple Sequence Alignment

Settings	Synthetic Datasets				Realistic Datasets	
Solvers\Data	Syn01 N=10, L=30 (3, 2, 4)	Syn02 N=30, L=50 (12, 11, 7)	Syn03 N=30, L=50 (19, 18, 9)	Syn04 N=30, L=50 (24, 24, 19)	sDicF N=6, L=15 (3, 4, 16)	sHairpin N=20, L=30 (9, 7, 44)
ClustalOmega	311 / 47	3295 / 126	6671 / 274	5946 / 240	119 / 27	1225 / 77
Kalign	88 / 10	1440 / 51	2003 / 71	2612 / 93	104 / 24	874 / 54
T-COFFEE	99 / 12	1031 / 36	1492 / 53	2120 / 75	104 / 24	868 / 53
MAFFET	87 / 10	1196 / 42	1856 / 66	2843 / 103	103 / 27	874 / 54
MUSCLE	87 / 10	1060 / 37	1649 / 59	2311 / 83	105 / 24	874 / 54
ConvexMSA	**79 / 9**	**863 / 30**	**1285 / 45**	**1903 / 67**	**98 / 23**	**853 / 50**
Ground Truth	*79 / 9*	*863 / 30*	*1310 / 46*	*1903 / 67*	*103 / 23*	*974 / 60*

- Synthetic data use TKF1 model (Thorne et al., 1991) to generate insertion/deletion (with some Poisson rate).
- (I,D,M)=(#insertions,#deletions,#mismatches).
- We report both Sum-of-Pairs / Star-Alignment scores.

Experimental Result: Motif Discovery

Text Encoded Case 1: $L_m = 4, L_M = 6$

01011000010100111000101111110101100010000100001000010
11111101011000100000110010

ConvexMD sol-1: matching rate 100.0%

010110 000101 001110 0010 111111 010110 0010 000100 0010
v *e* *n* *i* *v* *i* *d* *i*
111111 010110 0010 000011 0010
v *i* *c* *i*

ConvexMD sol-2: matching rate 100.0%

0101 1000 0101 0011 100010 111111 0101 1000 1000 0100 0010
1 2 1 3 4 5 1 2 2 6 7
111111 0101 1000 1000 0011 0010
5 1 2 2 3 7

MEME sol: matching rate 75.7%

0101 100001 01 0011 1 0001 0 111111 0 1011 000 100001 0 0001 0
7 3 4 6 1 2 3 6
111111 0 1011 0 0010 00 0011 0010
1 2 5 4

- Dechier is a Motif Discovery problem with perfect matched solution.
- Each character of a well-known saying vevi vidi vici (I came, I saw, I conquered) is encoded into a binary string and concatenated with others.
- We compare our convex relaxation approach to the current most-widely-used algorithm Multiple EM for Motif Elicitation (MEME) for MD.

Conclusion

- Multiple Sequence Alignment (MSA) and Motif Discovery (MD) are two fundamental and NP-hard tasks in Bioinformatics.
- We propose a convex relaxation approach to MSA & MD based on the concept of Atomic Norm, and a GDMM algorithm that can find its solution in practice.
- Experiments on small-scale MSA, MD problems demonstrate superioty of the convex approach. Scalability is one of our on-going works.