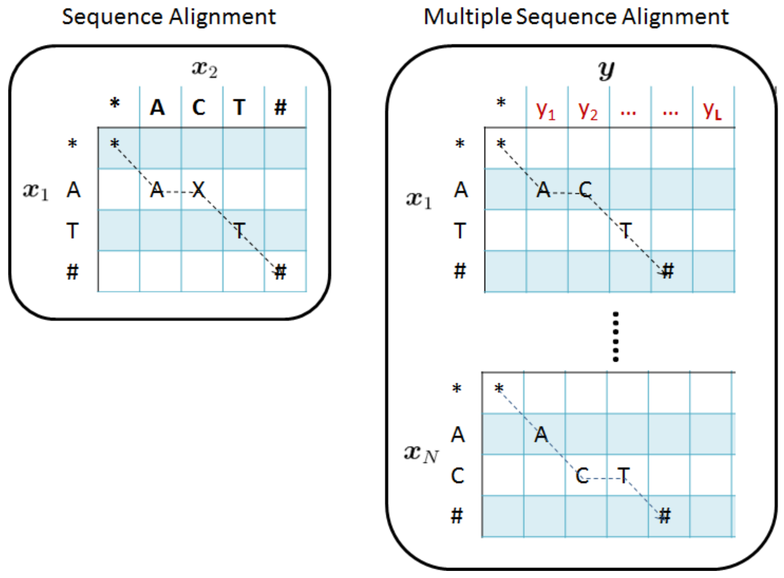
A Convex Atomic-Norm Approach to Multiple Sequence Alignment & Motif Discovery

Ian E.H. Yen*, Xin Lin*, Jiong Zhang, Pradeep Ravikumar, Inderjit S. Dhillon

Abstract

- Multiple Sequence Alignment (MSA) and Motif Discovery (MD) are two fundamental tasks in Bioinformatics that are known as NP-hard problems.
- Existing approaches are based on either local search methods such as Expectation Maximization (EM), Gibbs Sampling or greedy heuristic methods.
- We propose a convex relaxation approach to MSA & MD based on the concept of Atomic Norm.
- A Greedy Direction Method of Multiplier (GDMM) algorithm is proposed to solve the convex program with two atomic norm constraints.

Multiple Sequence Alignment & Motif Discovery



- **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 x_1, 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 (MSA)**:

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

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

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

  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: Reduce MSA to a series of greedily-selected Pairwise SA.
- Iterative, hill-climbing methods.

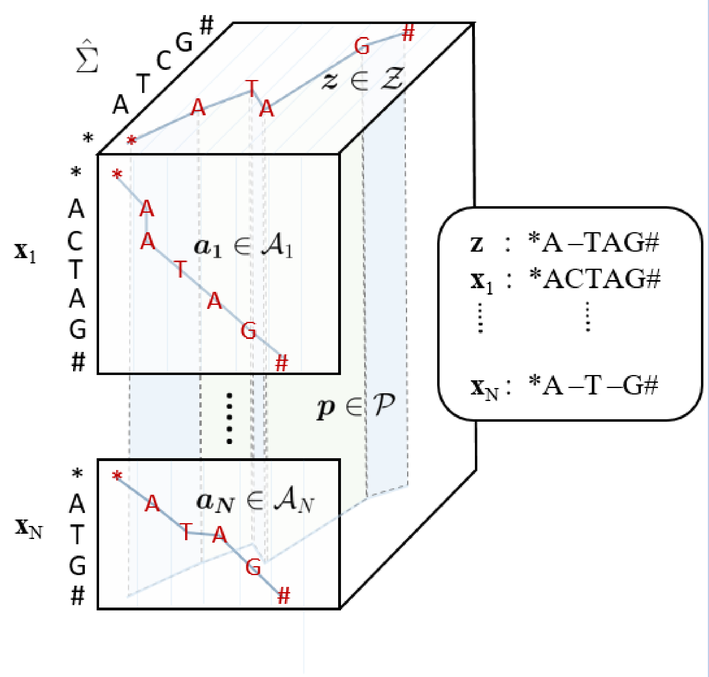
Convex Relaxation via Atomic Constraints

- $\mathcal{M} = \ell \times L \times |\hat{\Sigma}|$.
- $\mathcal{A}_n$: All possible aligns a_n.
- $\mathcal{P}$: Any "folding screen" p of single consensus z.
- MSA via Atomic Sets:

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

- Convex Relaxation:

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



Convex Relaxation of Motif Discovery

- Use similar representation 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 $\Omega_{\mathcal{P}}(W) \leq K$, where $\Omega_{\mathcal{P}}(W) := inf\{q \geq 0 : W \in q * Conv(\mathcal{P})\}$.

Greedy Direction Method of Multiplier (GDMM)

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

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

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

$$Y^{(t+1)} = Y^{(t)} + \eta \left(W_1^{(t+1)} - W_2^{(t+1)} \right), \quad (3)$$

  where Y is dual variable corresponding to constraint $W_1 = W_2$.
- However, the atomic constraints involve exponential number of atoms, so one can hardly minimize AL w.r.t. W_1 or W_2 exactly.
- We propose a GDMM algorithm, which minimizes (W_1, W_2) via only a (non-drop) step of Away-step Frank-Wolfe before each Dual Ascent (3).
- The Frank-Wolfe step only requires computation of the greedy atoms:

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

  and

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

  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.

Experiment: Multiple Sequence Alignment

- Synthetic uses 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.

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*

Experiment: Motif Discovery

- 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.