CONVEX:
De novo Transcriptome Error Correction by Convexification

Meisam Razaviyayn
Stanford University

Jeremy Guo
Stanford University

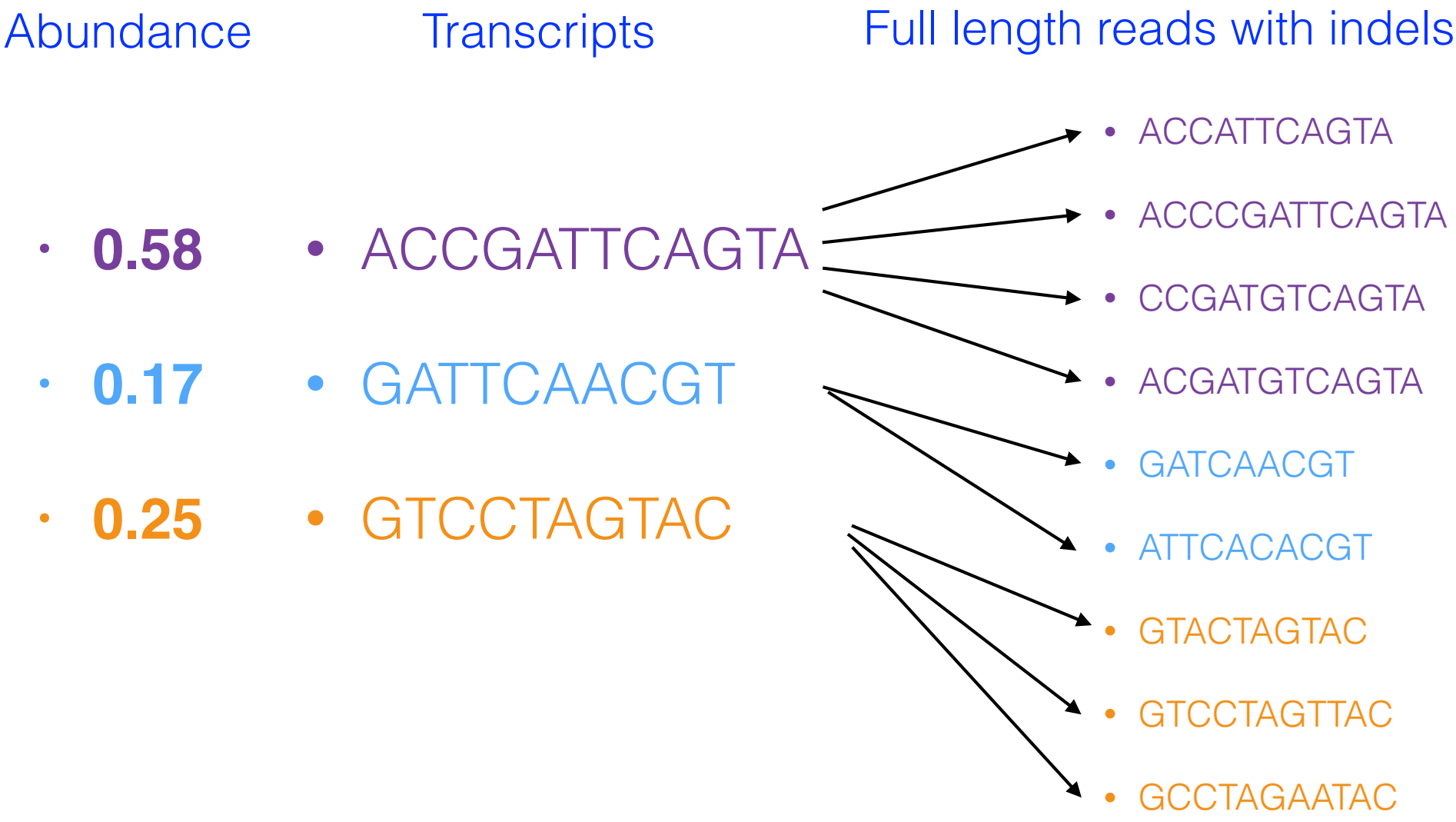
David Tse
Stanford University - U.C. Berkeley

Elizabeth Tseng
Pacific Biosciences

Problem Statement

Abundance	Transcripts
0.58	ACCGATTCAGTA
0.17	GATTCAACGT
0.25	GTCCTAGTAC

Problem Statement



Problem Statement

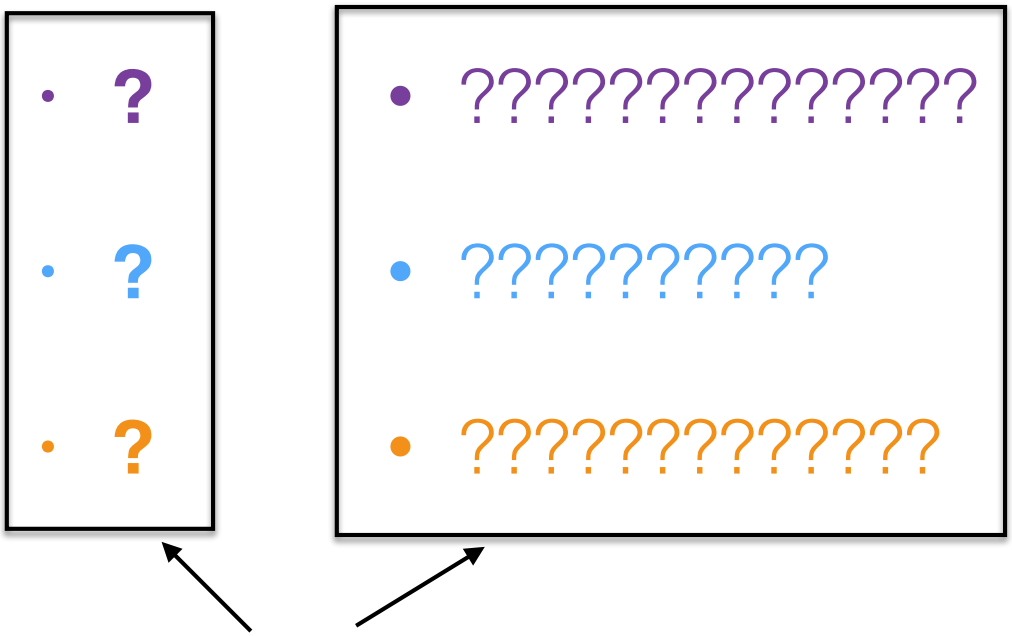
Abundance	Transcripts	Full length reads with indels
• ?	• ????????????	• ACCATTCAGTA
• ?	• ?????????	• ACCCGATTCAGTA
• ?	• ????????????	• CCGATGTCAGTA
		• ACGATGTCAGTA
		• GATCAACGT
		• ATTCACACGT
		• GTACTAGTAC
		• GTCCTAGTTAC
		• GCCTAGAATAC

Convexification

Abundance

Transcripts

Full length reads with indels



Maximum Likelihood Estimation

Nonconvex and NP-hard

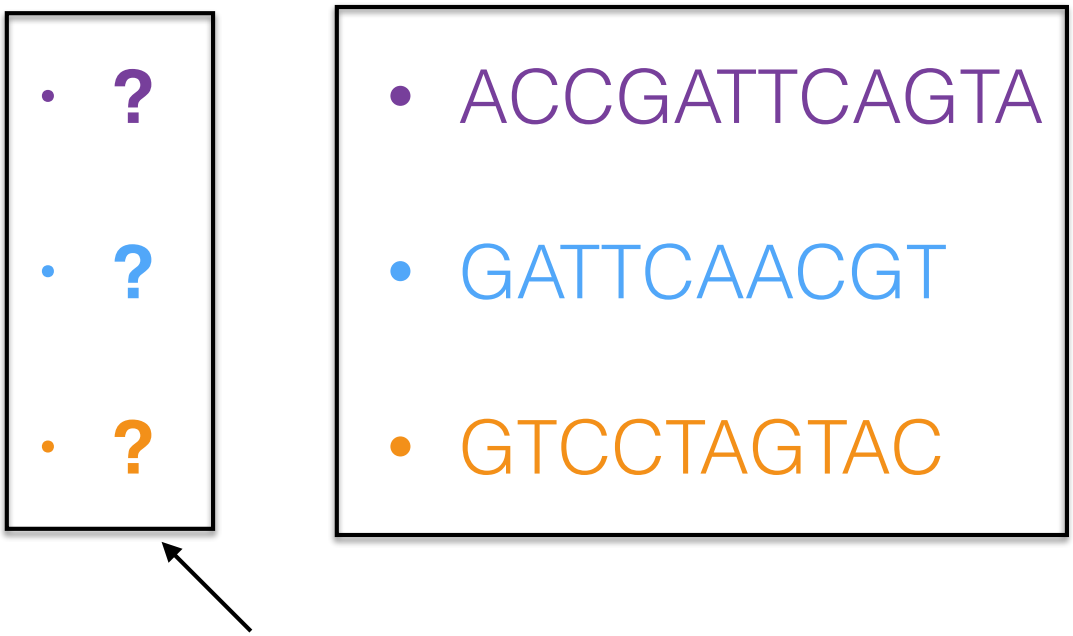
- ACCATTCAGTA
- ACCCGATTCAGTA
- CCGATGTCAGTA
- ACGATGTCAGTA
- GATCAACGT
- ATTCACACGT
- GTACTAGTAC
- GTCCTAGTTAC
- GCCTAGAATAC

Convexification

Abundance

Transcripts

Full length reads with indels



Maximum Likelihood Estimation

Convex and Easy

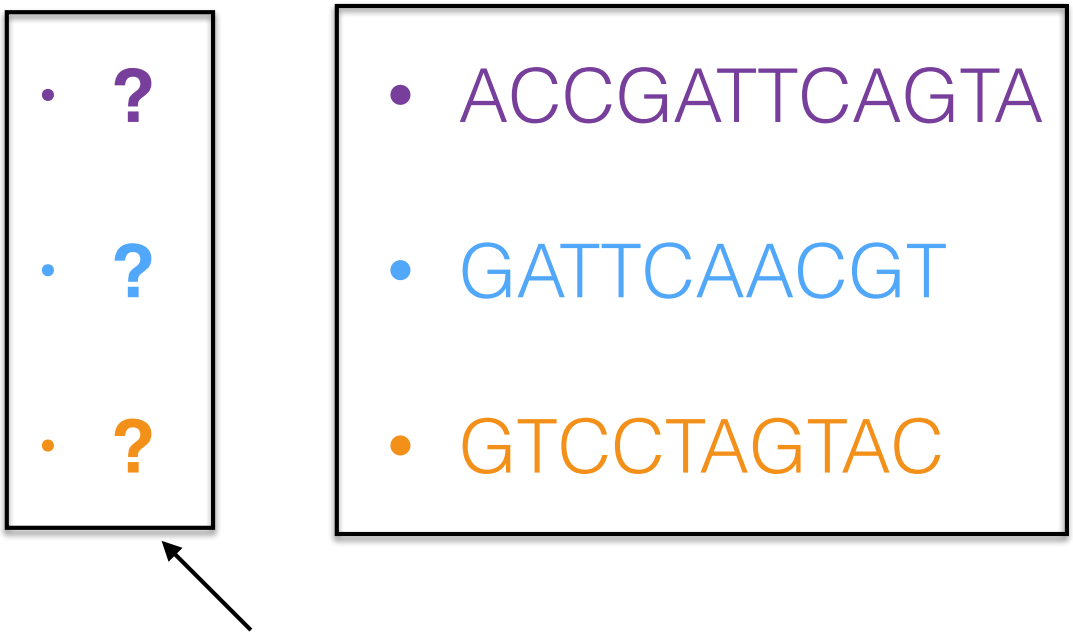
- ACCATTCAGTA
- ACCCGATTCAGTA
- CCGATGTCAGTA
- ACGATGTCAGTA
- GATCAACGT
- ATTCACACGT
- GTACTAGTAC
- GTCCTAGTTAC
- GCCTAGAATAC

Convexification

Abundance

Transcripts

Full length reads with indels



Maximum Likelihood Estimation

Convex and Easy

- ACCATTCAGTA
- ACCCGATTCAGTA
- CCGATGTCAGTA
- ACGATGTCAGTA
- GATCAACGT
- ATTCACACGT
- GTACTAGTAC
- GTCCTAGTTAC
- GCCTAGAATAC

Set of all possible sequences? **Convex but exponentially many**

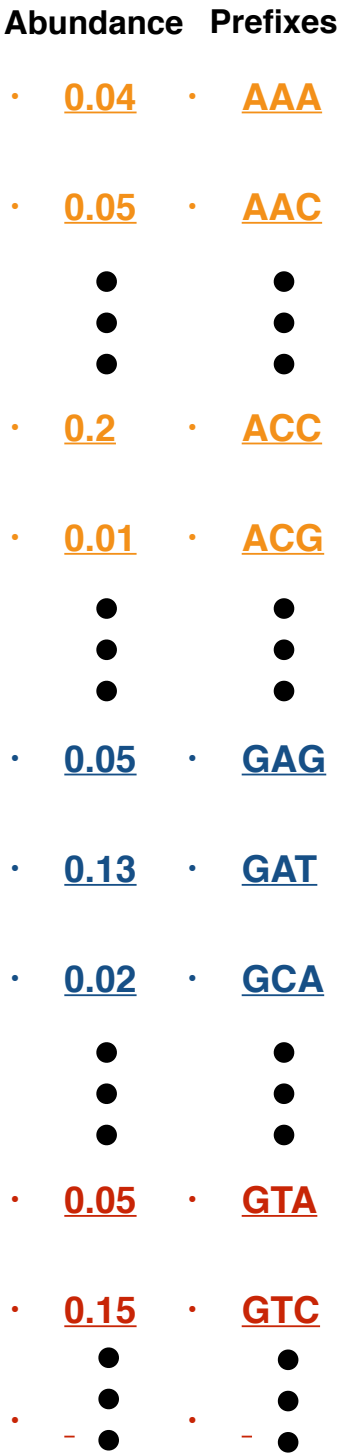
Greedy Algorithm

Prefixes

- AAA
- AAC
- ACC
- ACG
- GAG
- GAT
- GCA
- GTA
- GTC

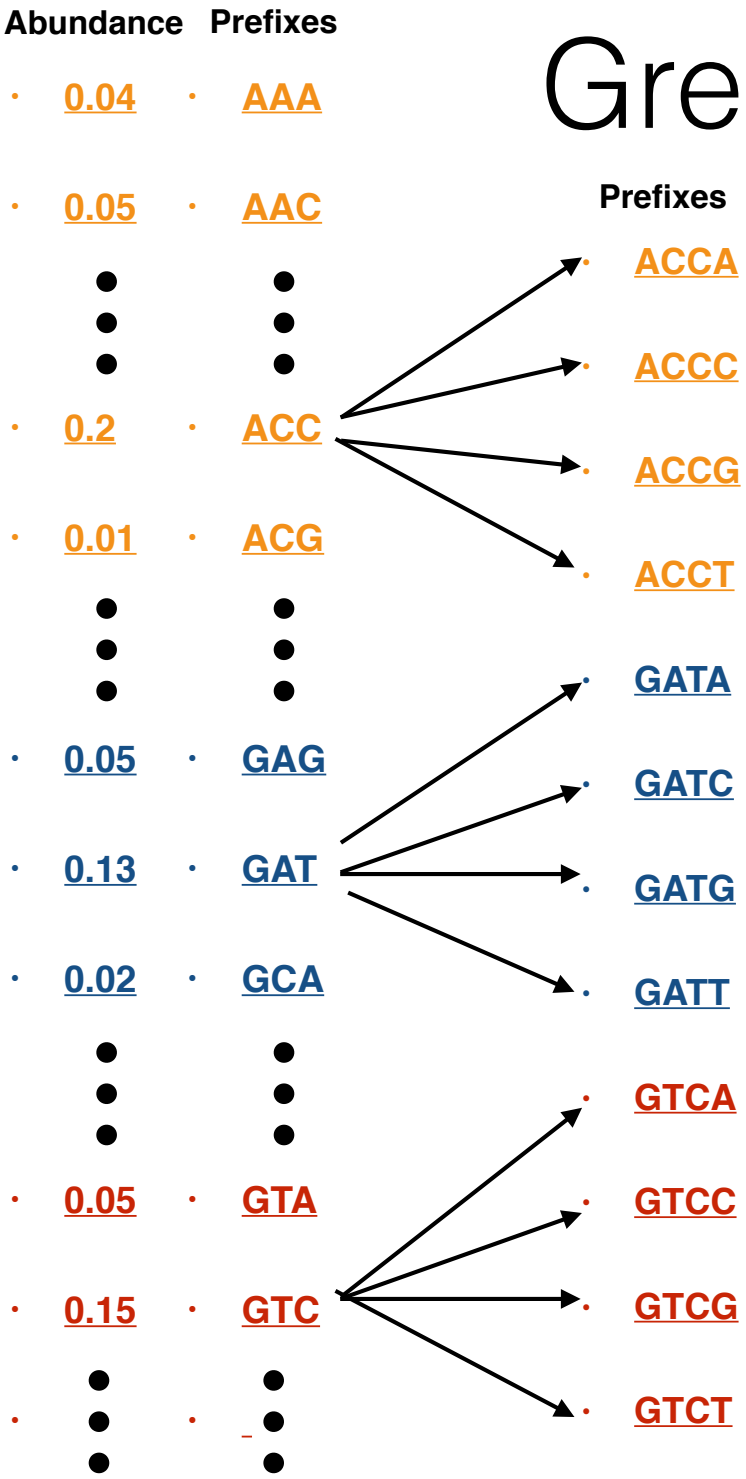
- ACCATTCAGTA
- ACCCGATTCAGTA
- CCGATGTCAGTA
- ACGATGTCAGTA
- GATCAACGT
- ATTCACACGT
- GACTAGTAC
- GCCTAGTTAC
- GACTAGAATAC

Greedy Algorithm



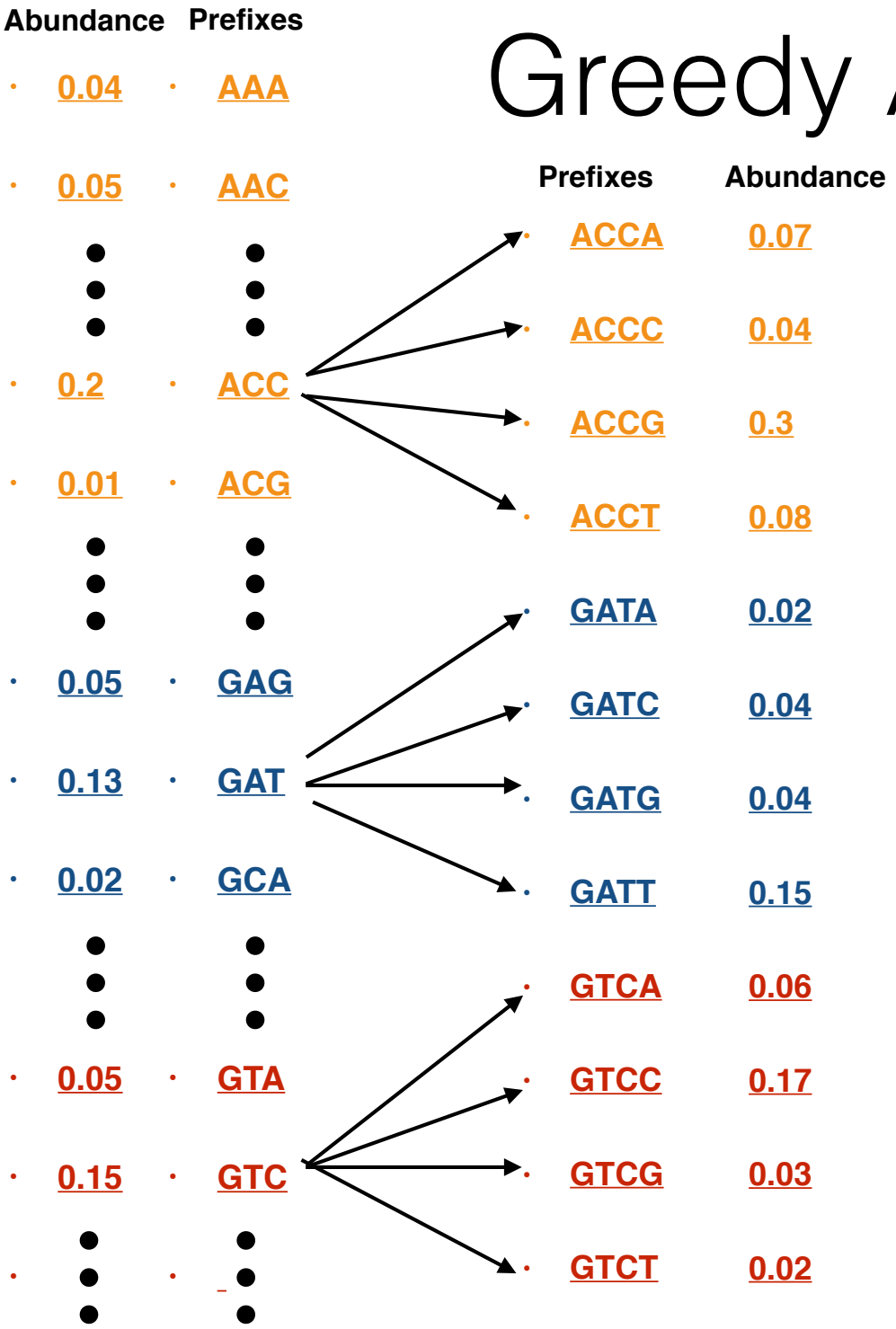
- ACCATTCAGTA
- ACCCGATTCAGTA
- CCGATGTCAGTA
- ACGATGTCAGTA
- GATCAACGT
- ATTCACACGT
- GACTAGTAC
- GCCTAGTTAC
- GACTAGAATAC

Greedy Algorithm



- ACCATTCAGTA
- ACCCGATTCAGTA
- CCGATGTCAGTA
- ACGATGTCAGTA
- GATCAACGT
- ATTCACACGT
- GACTAGTAC
- GCCTAGTTAC
- GACTAGAATAC

Greedy Algorithm



- ACCATTCAGTA
- ACCCGATTCAGTA
- CCGATGTCAGTA
- ACGATGTCAGTA
- GATCAACGT
- ATTCACACGT
- GACTAGTAC
- GCCTAGTTAC
- GACTAGAATAC

Greedy Algorithm

Prefixes	Abundance
ACCA	0.07
ACCC	0.04
ACCG	0.3
ACCT	0.08
GATA	0.02
GATC	0.04
GATG	0.04
GATT	0.15
GTCA	0.06
GTCC	0.17
GTCG	0.03
GTCT	0.02

- ACCATTCAGTA
- ACCCGATTCAGTA
- CCGATGTCAGTA
- ACGATGTCAGTA
- GATCAACGT
- ATTCACACGT
- GACTAGTAC
- GCCTAGTTAC
- GACTAGAATAC

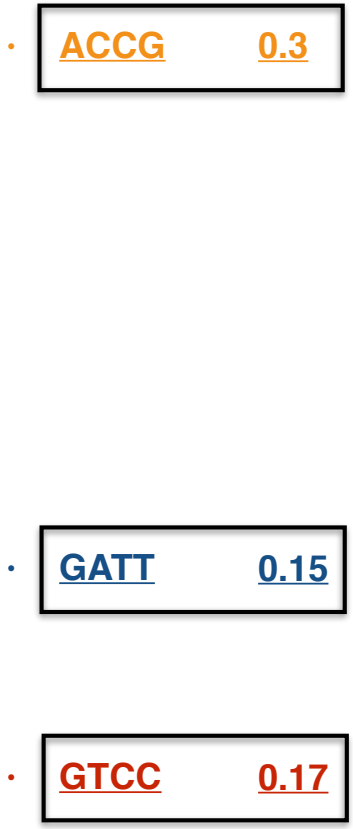
Greedy Algorithm

Prefixes	Abundance
ACCA	0.07
ACCC	0.04
ACCG	0.3
ACCT	0.08
GATA	0.02
GATC	0.04
GATG	0.04
GATT	0.15
GTCA	0.06
GTCC	0.17
GTCG	0.03
GTCT	0.02

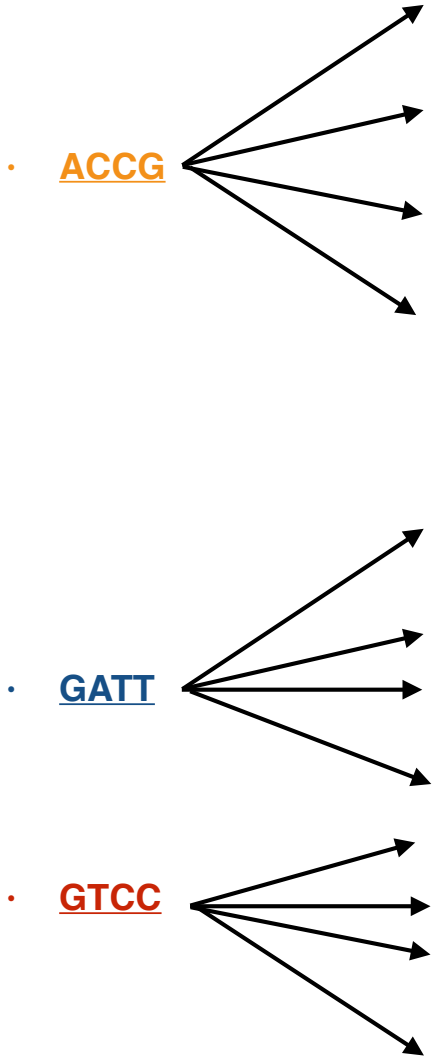
- ACCATTCAGTA
- ACCCGATTCAGTA
- CCGATGTCAGTA
- ACGATGTCAGTA
- GATCAACGT
- ATTCACACGT
- GACTAGTAC
- GCCTAGTTAC
- GACTAGAATAC

Greedy Algorithm

- ACCATTCAGTA
- ACCCGATTCAGTA
- CCGATGTCAGTA
- ACGATGTCAGTA
- GATCAACGT
- ATTCACACGT
- GACTAGTAC
- GCCTAGTTAC
- GACTAGAATAC



Greedy Algorithm



- ACCATTCAGTA
- ACCCGATTCAGTA
- CCGATGTCAGTA
- ACGATGTCAGTA
- GATCAACGT
- ATTCACACGT
- GACTAGTAC
- GCCTAGTTAC
- GACTAGAATAC

Remarks

- Consistency
-
-

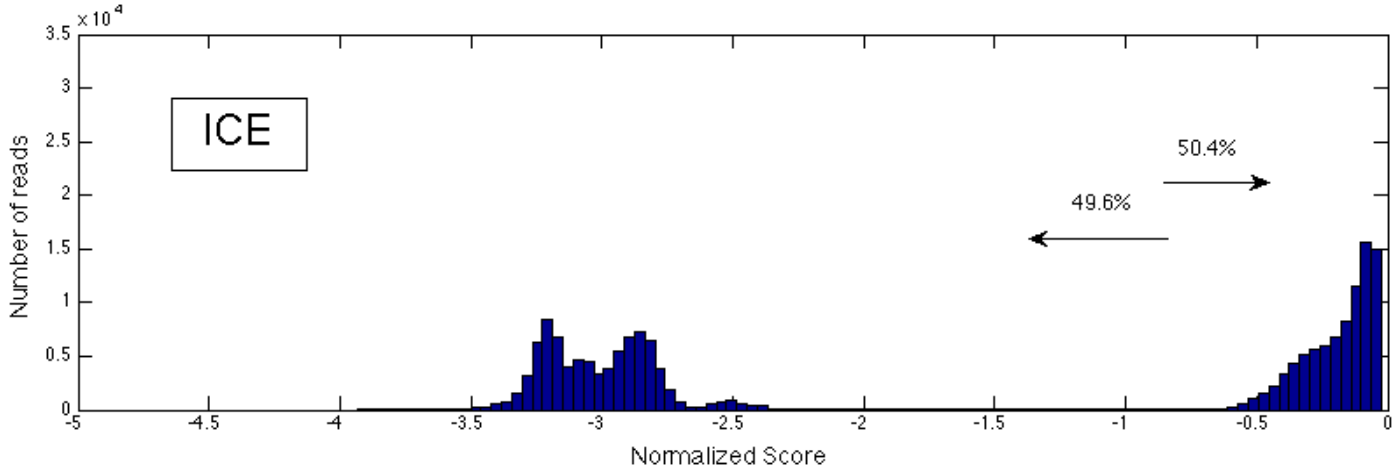
Remarks

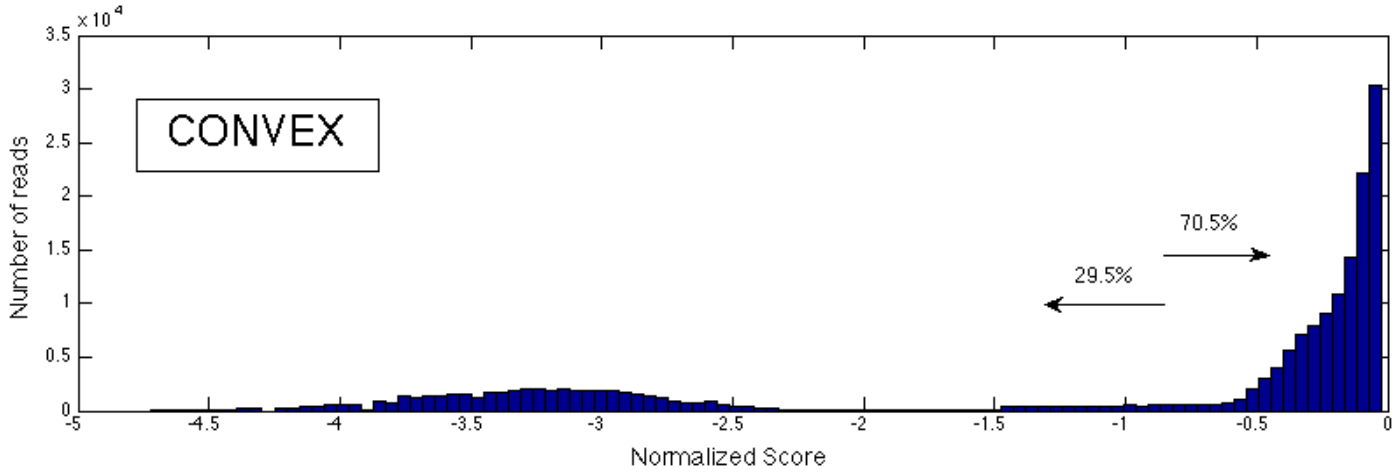
- Consistency
- No pairwise alignment between reads
-

Remarks

- Consistency
- No pairwise alignment between reads
- Linear computational complexity

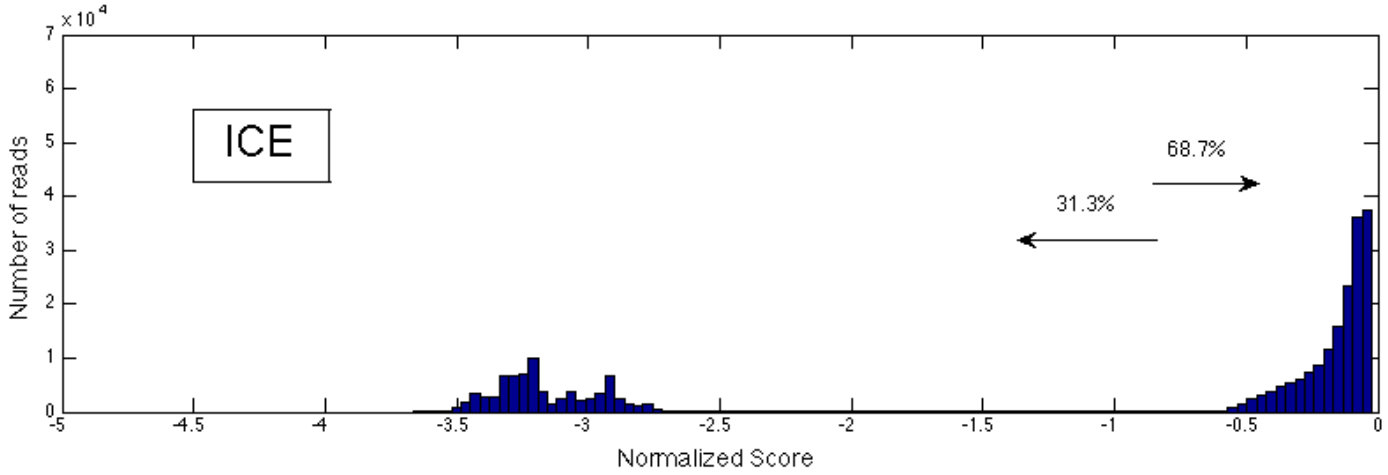
Numerical Experiments: Heart Tissue

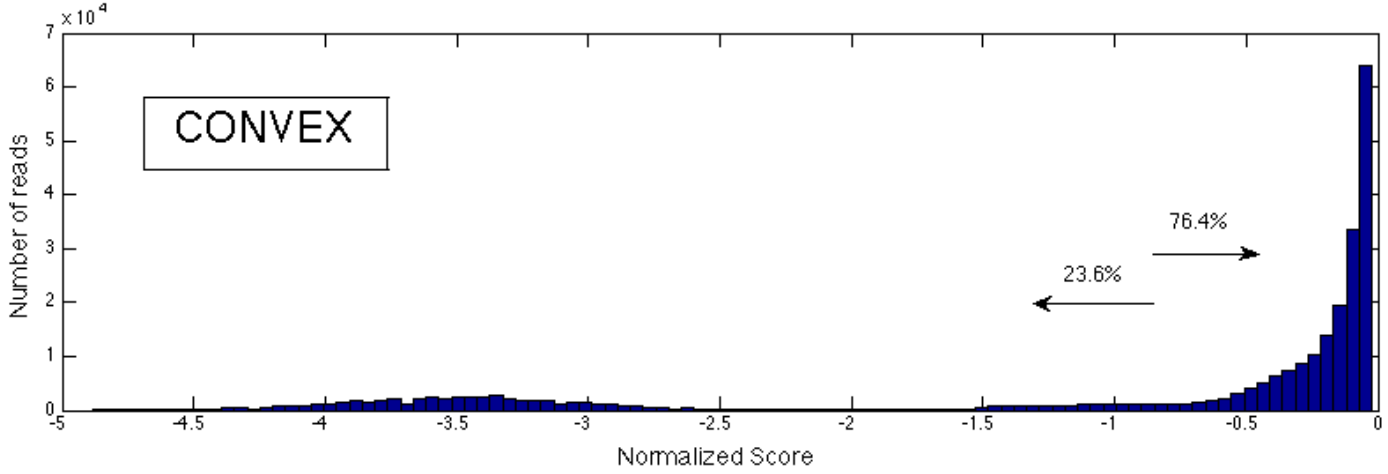




Number of CCS FL Reads	173,566
Number of Transcripts	6,896
Average Length	2,761

Numerical Experiments: Liver Tissue





Number of CCS FL Reads	245,381
Number of Transcripts	6,124
Average Length	2,198

Questions?

Estimating Abundances: Mixture Model

- Isoforms/centroids $s_1, s_2, \ldots, s_M$
- Abundances $\rho_1, \rho_2, \ldots, \rho_M$

- Full length reads $r_1, r_2, \ldots, r_N$
- Mixture model: $$\mathbb{P}\left(r_n; \mathbf{s}, \boldsymbol{\rho}\right) = \sum_{m=1}^{M} \rho_m \mathbb{P}\left(r_n \mid r_n \text{ from } s_m\right)$$

- Maximum likelihood estimation: $$\arg\max_{\boldsymbol{\rho}} \sum_{n=1}^{N} \log\left(\sum_{m=1}^{M} \rho_m \mathbb{P}\left(r_n \mid r_n \text{ from } s_m\right)\right)$$