Fast Motif Discovery in Short Sequences

Honglei Liu[1], Fangqiu Han[1], Hongjun Zhou[2], Xifeng Yan[1], Kenneth S. Kosik[2]

[1]Department of Computer Science, , University of California, Santa Barbara

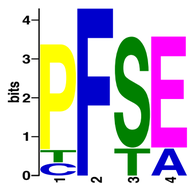
[2]Neuroscience Research Institute, University of California, Santa Barbara

Introduction

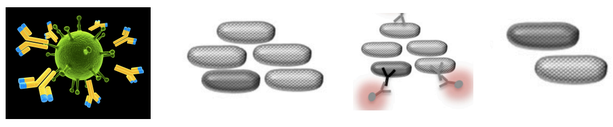
- **Motif**: frequently appearing sequence patterns
- **Motif discovery:**

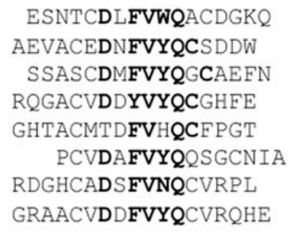
consensus string: PFSE

PWM (Position Weight Matrix)



- **Applications**
  - Transcription factor binding sites (TFBSs) discovery
  - Antibody biomarkers discovery





Large scale, Large alphabet set, Short

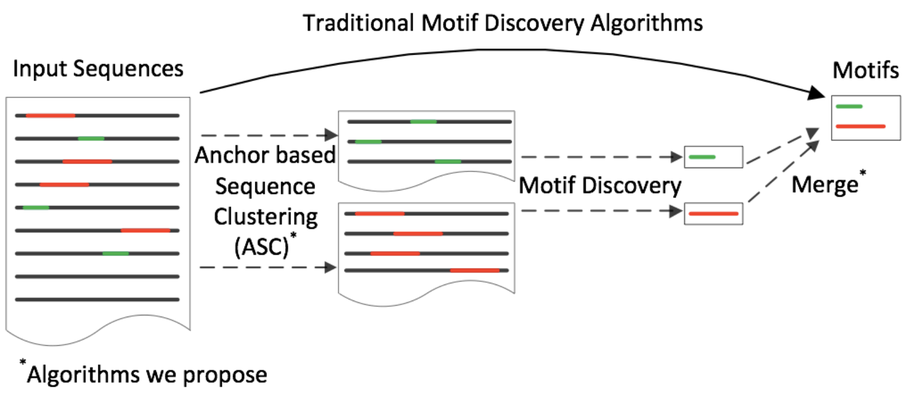
Challenges

- Unknown: number of motifs, length of motifs, etc
- Before next-generation sequencing era
  - At most several hundred sequences
- After next-generation sequencing era
  - Tens of thousands or even millions of sequences
- Existing methods can not handle the big data challenge very well

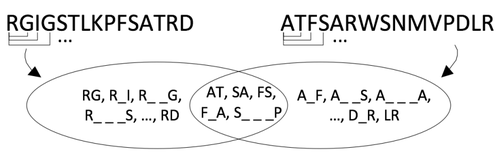
	Can handle >10k seq.	Can handle >1M seq.	Can work with protein seq.	Accuracy is as good as MEME
MEME[Bailey94,06]			✔	✔
STEME[Reid11]	✔	✔		✔
DREME[Bailey11]	✔	✔		
GibbsCluster[Andreatta13]	✔		✔	
MUSI[Kim11]	✔		✔	
Our framework	✔	✔	✔	✔

Methods

- **Our framework**
  - Reuse existing techniques!



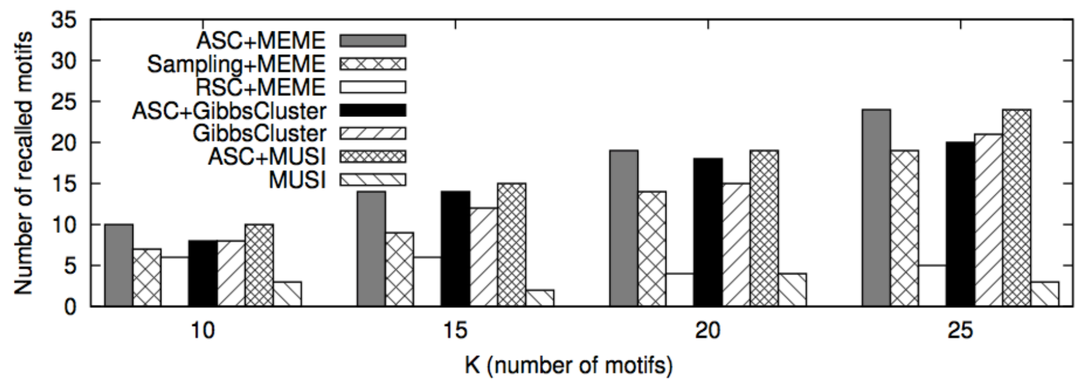
- **Anchor based Sequence Clustering algorithm (ASC)**
  - Could capture local similarities
  - Avoid pairwise comparisons
- **Anchor based similarity**
  - Represent sequences as *q*-anchor sets
  - e.g. 2-anchors of *PFSE* are {*PF, FS, SE, P_S, F_E, P_ _E*}



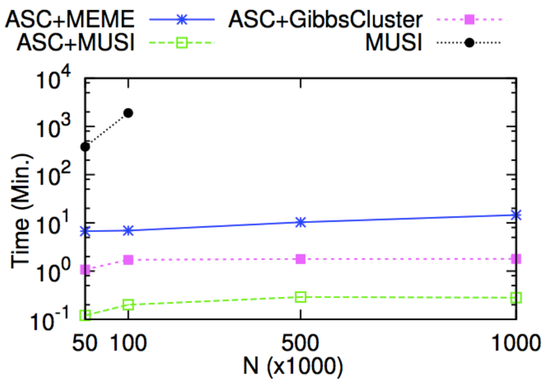
- **Iterative process**
  - Choose initial centers using *odd score*
    - Indicates how likely an anchor is from a motif
  - Adjust centers using *abundance score*
    - Indicates how unique an anchor is for a motif

Experiments

- Real data shows that our framework can reduce the runtime of MEME from **weeks** to **minutes** without losing accuracy!
- Apply ASC on top of MEME, MUSI and GibbsCluster
- Number of recalled motifs from different methods using synthetic data (10k seq.)



- Scalability



Conclusions

- Big data challenge
- Reuse existing techniques
- Huge performance gain without losing accuracy