Min-Sum Clustering of Protein Sequences with Limited Distance Information

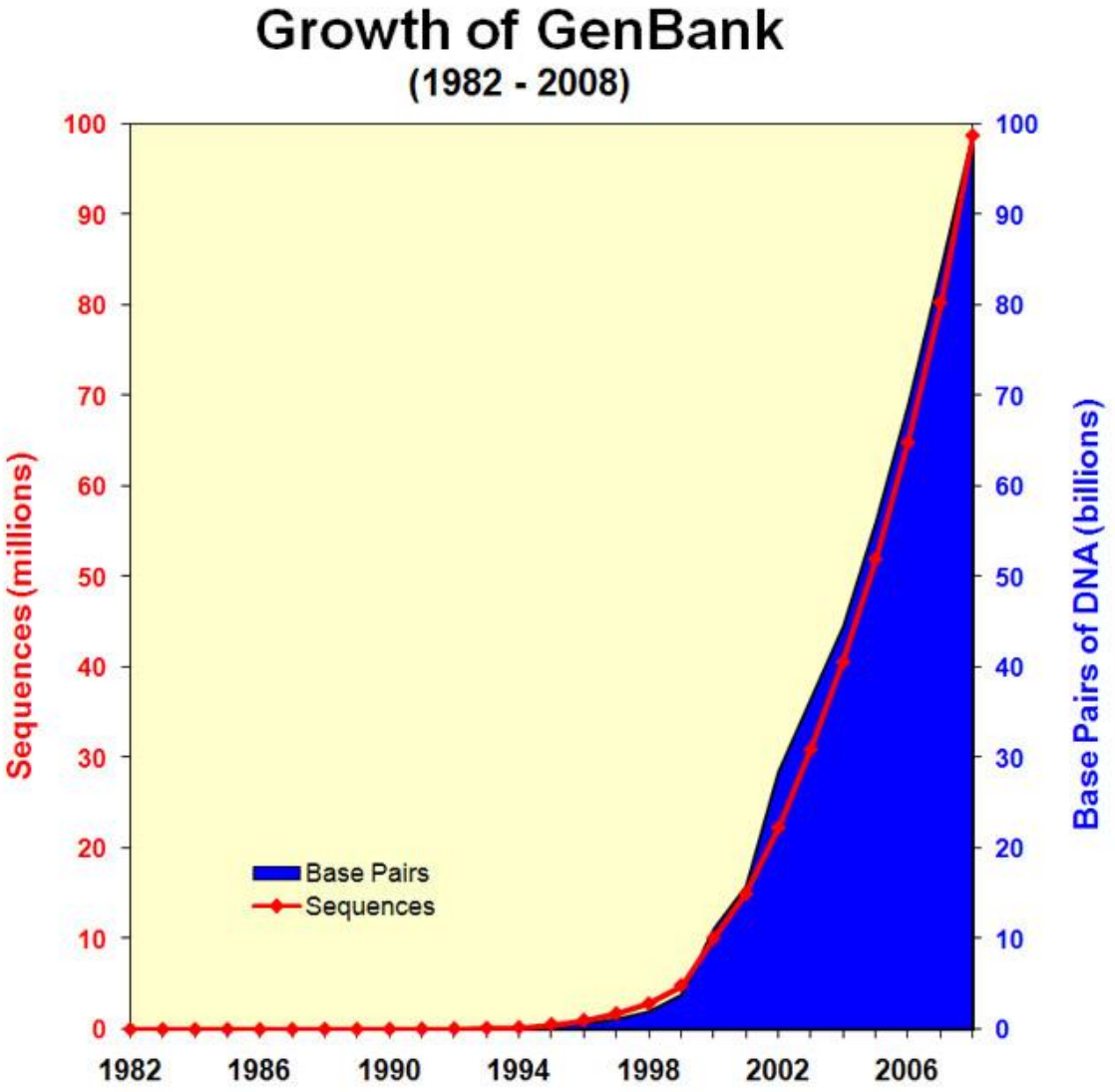
Konstantin Voevodski, Maria-Florina Balcan, Heiko Roglin, Shang-Hua Teng, Yu Xia

Outline

- Motivation
- Clustering Accuracy
- Algorithm Overview and Analysis
- Computational Experiments

Motivation

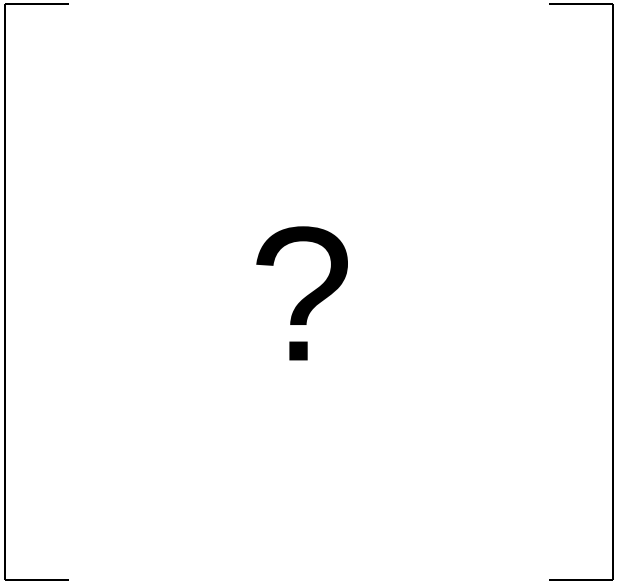
- Classifying proteins using sequence similarity is a very well-studied problem in computational biology.
- Size of biological sequence databases is rapidly expanding.



“from 1982 to the present, the number of bases in GenBank has doubled approximately every 18 months” (from release notes for release 162.0, October 2007)

Clustering with limited information

- Sequence comparison is computationally intensive.
- Computing all pairwise similarities may take orders of magnitude more time than performing the actual clustering.



Clustering with Limited Information

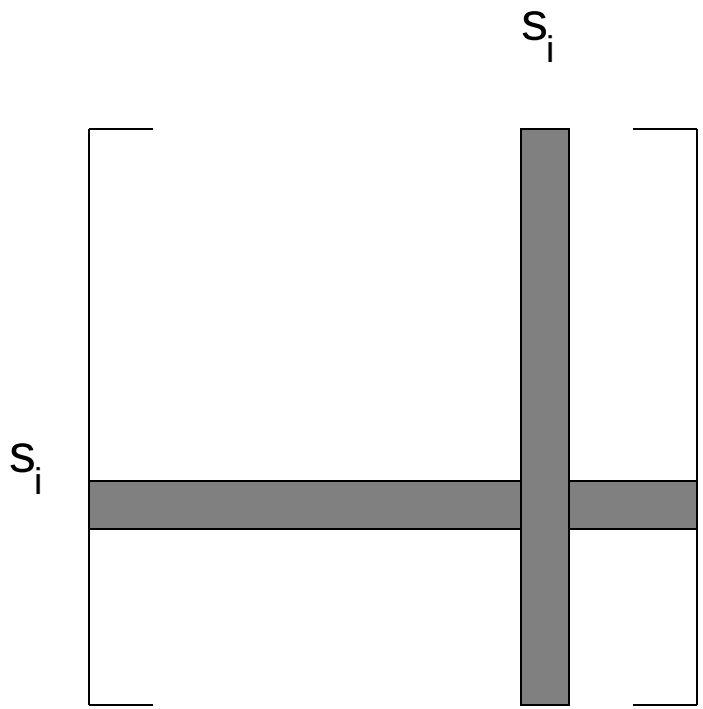
Limited information setting: algorithm queries some distances between objects as Needed during execution.

Queries are expensive.

Task: find accurate clustering using few queries.

Clustering with Limited Information

- Assume have access to one versus all queries.



One versus all distance queries

- Especially relevant for sequence similarity search in Biology.
- BLAST (Basic Local Alignment Search Tool) is optimized to search a single sequence against an entire database.
- *n* pairwise alignments take much longer.
- BLAST still gives meaningful results.

Outline

- Motivation
- Clustering Accuracy
- Algorithm Overview and Analysis
- Computational Experiments

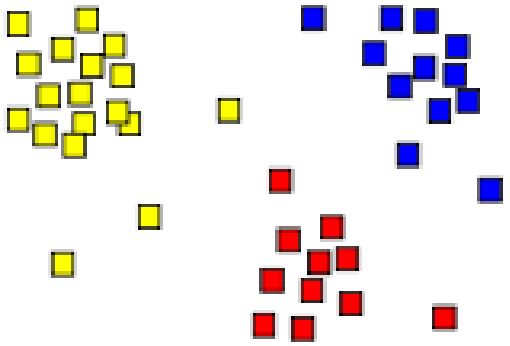
Accuracy

- Assume approximation stability of min-sum objective function for clustering.
- Approximation stability property of Balcan, Blum, Gupta (2009): clusterings that approximate the objective function well are structurally close to “target” clustering.
- Task: find clustering close to the target.

Objective functions for clustering

Would like to find a k-clustering C that partitions the points in S into k sets $C_1, C_2, \ldots C_k$.

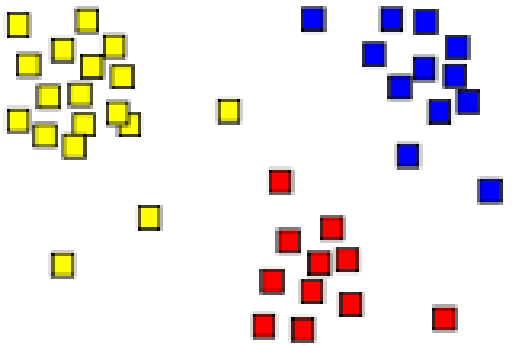
Can define some objective function to measure quality of clustering, then optimize it.



Objective functions for clustering

min-sum objective: minimize $\Phi(C) = \sum_{i=1}^{k} \sum_{x,y \in C_i} d(x,y)$.

Denote by OPT_Φ its optimal value: $OPT_\Phi = \min_C \Phi(C)$, where the minimum is over all k-clusterings of S.



Approximation stability property

Any clustering C that approximates OPT_Φ within a factor of c is ϵ-close to C_T:

$$\Phi(C) \leq c \cdot OPT_\Phi \Rightarrow dist(C, C_T) < \epsilon.$$

The distance between two clusterings C and C' is the fraction of mismatched points under the optimal matching between the two sets of clusters:

$$dist(C, C') = \min_{\sigma \in S_k} \frac{1}{n} \sum_{i=1}^{k} |C_i - C'_{\sigma(i)}|,$$

where S_k is the set of bijections $\sigma\colon [k] \to [k]$.

Approximation stability property

- [Balcan, Blum, Gupta]: Given the (c, ϵ)-property can efficiently find clusterings structurally close to the target without approximating the objective function.
- True even when c-approximation of the objective function is hard to compute.

Clustering with limited information

- Possible to find accurate clusterings in this model even without knowing most of the distances between the points.
- Considering only some of the distances also gives more efficient algorithms.

Outline

- Motivation
- Clustering Accuracy
- Algorithm Overview and Analysis
- Computational Experiments

Algorithm Overview

- Select a small number of points (landmarks).
- Only use distances between landmarks and other points.

Algorithm Description

Landmark-Clustering-Min-Sum(S, d, k, n', T):

- Choose a set of landmarks L of size n' uniformly at random from S
- $i = 1$, $r = 0$
- WHILE $i \leq k$
  - for each $l \in L$: $B_l = \{s \in S \mid d(s, l) \leq r\}$
  - if $\exists l^* \in L : |B_{l^*}| \cdot r > T$
    - $L' = \{l \in L : \ B_l \cap B_{l^*} \neq \emptyset\}$
    - $C_i = \{s \in S : \ s \in B_l \text{ and } l \in L'\}$
    - remove points in C_i from consideration
    - $i = i + 1$
  - increment r to the next relevant distance
- RETURN $C = \{C_1, \ldots C_k\}$

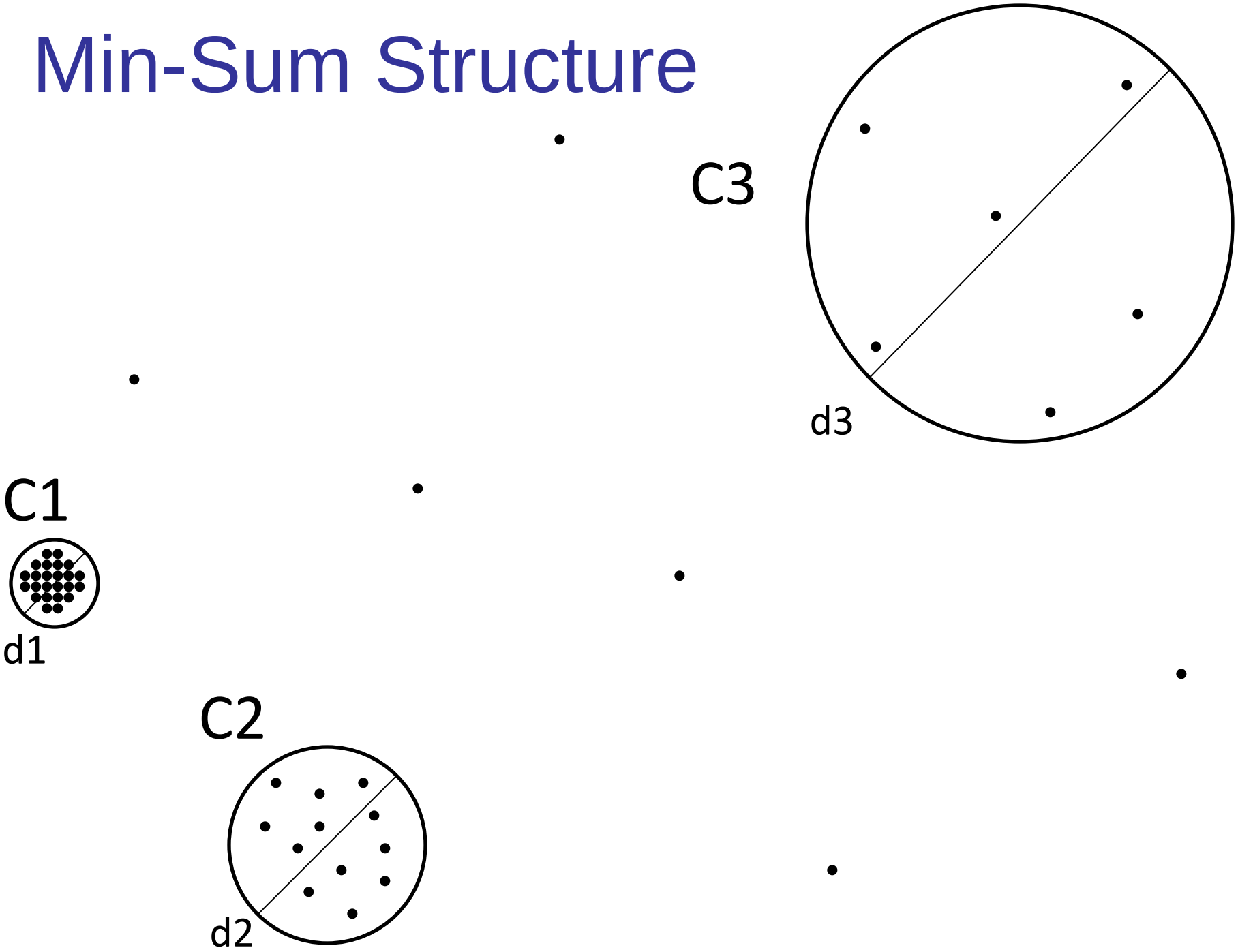
Theoretic Results

If the instance satisfies the $(2(1+\alpha),\epsilon)$-property for the min-sum objective, we are given the optimum objective value, and each cluster in the target clustering is large, then with high probability **Landmark-Clustering-Min-Sum** finds a clustering that is $\mathrm{O}(\epsilon/\alpha)$-close to C_T in time *O*(*k* log (*k*) *n* log (*n*)) using *O*(*k* log *k*) **one versus all** distance queries.

Proof Outline

- If approximation stability holds for the min-sum objective, the data must have a certain structure.
- Each target cluster must contain a core of well-separated points.
- Can find this clustering as long as we get a landmark in each cluster core.

Min-Sum Structure



Theoretic Results

- If do not know OPT_Φ must run algorithm with increasing estimate of *T* until enough points are clustered.
- Need $|L| \cdot n^2$ iterations to provably find an accurate clustering.
- In practice the number of iterations is much smaller.

Outline

- Motivation
- Clustering Accuracy
- Algorithm Overview and Analysis
- Computational Experiments

Computational experiments

- Cluster proteins by sequence similarity.
- BLAST is one-versus-all distance query.
- Compare with “gold standard” manual classifications given in the Pfam database.
- Use distance measure from theoretic part of our work to compare clusterings.

Computational Experiments

- Datasets chosen by randomly selecting families from pFam.
- Datasets are too large to compute the full distance matrix.
- Can only compare with a method that uses a similar amount of distance information.

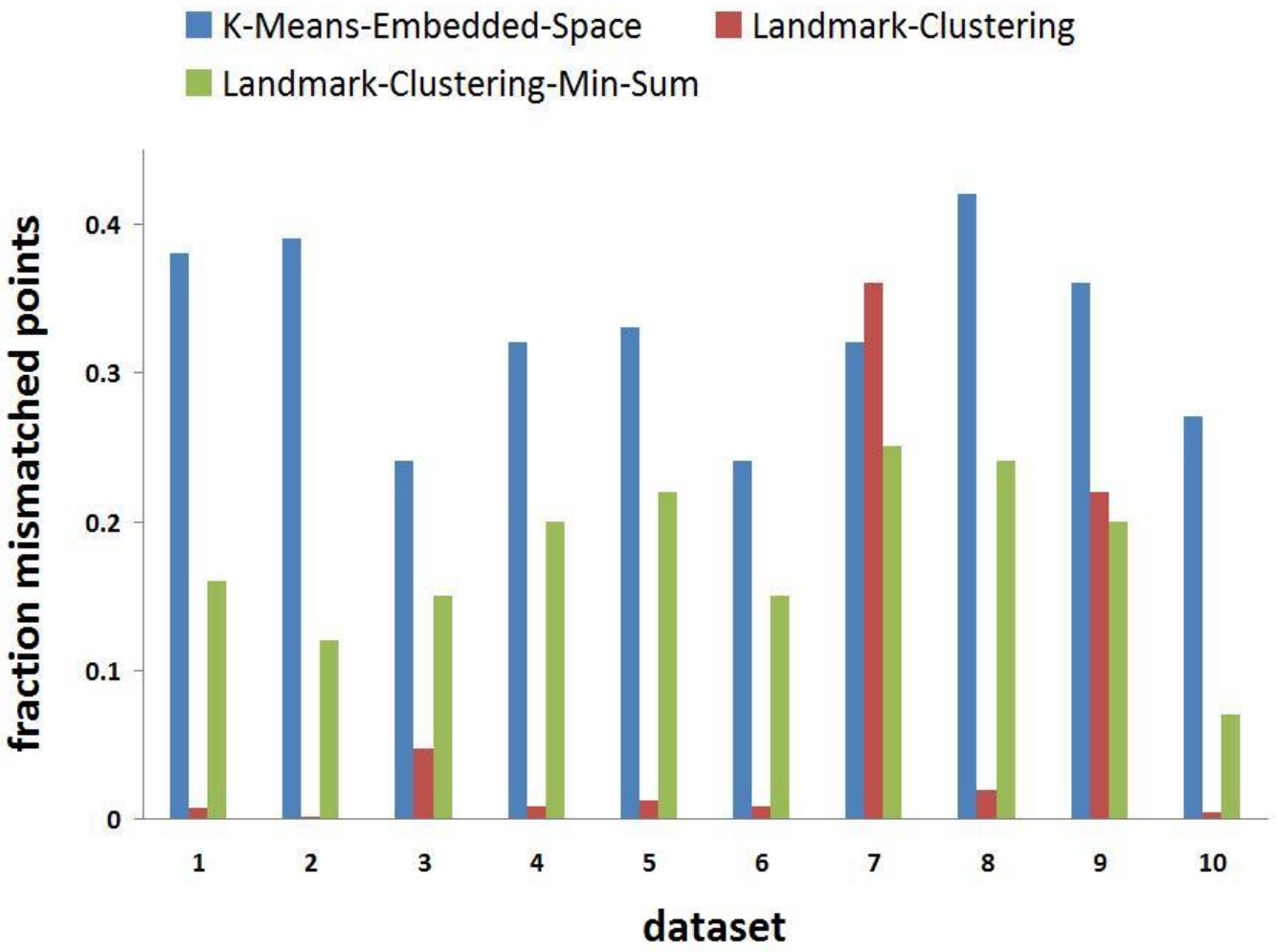
Computational Experiments

- The alternative method selects a set of landmarks, builds a coordinate system using them, and performs *k*-means clustering in this space:

  K-Means-Embedded-Space(*d,k*):

  - Randomly choose a set of landmarks L, $|L| = d$
  - Embed each point in a *d*-dimensional space using distance to L
  - Use *k*-means clustering in this space

- Also compare with our other limited-information clustering algorithm (from previous work).



Comparing the performance of ***K*-Means-Embedded-Space** (blue), **Landmark-Clustering** (red), and **Landmark-Clustering-Min-Sum** (green) on 10 datasets from Pfam. Datasets **1-10** are created by randomly choosing 8 families from Pfam of size *s*, 1000 <= *s* <= 10000.

Computational Experiments

- Landmark-Clustering finds accurate clustering given approximation stability of *k-median* objective function (Voevodski, Balcan, Roglin, Teng, Xia, 2009).
- Pfam sequence data resembles *k-median* structure.

Future Directions

- Find data that has (non-trivial) min-sum structure: clusters of varying size, diameters inversely proportional to size.
- Smarter landmark selection: adaptively choose landmarks (like the *k*-median algorithm).
- Limited information clustering given other assumptions about structure of ground truth.

Outline

- Motivation
- Clustering Accuracy
- Algorithm Overview and Analysis
- Computational Experiments

Thank you!!