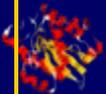


# High-throughput Biological Data

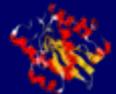
- Enormous amount of biological data are being generated by high-throughput capabilities; even more are coming
  - ↳ genomic sequences
  - ↳ gene expression data
  - ↳ mass spec. data
  - ↳ protein-protein interaction
  - ↳ protein structures
  - ↳ .....
- Hidden in these data is information that reflects
  - ↳ existence, organization, activity, functionality ..... of biological machineries at different levels in living organisms

Most effectively utilizing this information will prove to be essential in successful implementation of GTL



# Data Issues .....

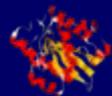
- Data collection: getting the data
- **Data representation**: data standards .....
- Data organization and storage: database issues .....
- **Data analysis and data mining**: discovering “knowledge”, **patterns/signals**, from data, establishing **associations** among data patterns
- Data utilization and application: from data patterns/signals to models for bio-machineries
- **Data visualization**: viewing complex data .....
- Data transmission: data collection, retrieval, .....
- .....
- .....



# Bio-Data Analysis and Data Mining

- Existing/emerging bio-data analysis and mining tools for
  - ↙ DNA sequence assembly
  - ↙ Genetic map construction
  - ↙ Sequence comparison and database search
  - ↙ Gene finding
  - ↙ ....
  - ↙ Gene expression data analysis
  - ↙ Phylogenetic tree analysis to infer horizontally-transferred genes
  - ↙ Mass spec. data analysis for protein complex characterization
  - ↙ .....
- Current mode of work

**Developing ad hoc tools for each individual application**



# Bio-Data Analysis and Data Mining

- As the amount and types of data and the needs to establish connections across multi-data sources increase rapidly, the number of analysis tools needed will go up “exponentially”
  - ↳ blast, blastp, blastx, blastn, ... from BLAST family of tools
  - ↳ gene finding tools for human, mouse, fly, rice, cyanobacteria, .....
  - ↳ tools for finding various signals in genomic sequences, protein-binding sites, splice junction sites, translation start sites, .....

**Many of these data analysis problems are fundamentally the same problem(s) and can be solved using the same set of tools**

**Developing ad hoc tools for each application problem (by each group of individual researchers) may soon become inadequate as bio-data production capabilities further ramp up**

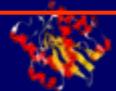


# Bio-data Analysis and Data Mining

---

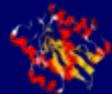
**To have analysis capabilities covering wide range of problems, we may have to discover the common fundamental structures of these problems**

**It is possible to develop a data analysis infrastructure in support of GTL and beyond**



# Data Clustering

- Many biological data analysis problems can be formulated as clustering problems
  - ↳ microarray gene expression data analysis
  - ↳ identification of regulatory binding sites (similarly, splice junction sites, translation start sites, .....
  - ↳ (yeast) two-hybrid data analysis (for inference of protein complexes)
  - ↳ phylogenetic tree clustering (for inference of horizontally transferred genes)
  - ↳ protein domain identification
  - ↳ identification of structural motifs
  - ↳ prediction reliability assessment of protein structures
  - ↳ NMR peak assignments
  - ↳ .....

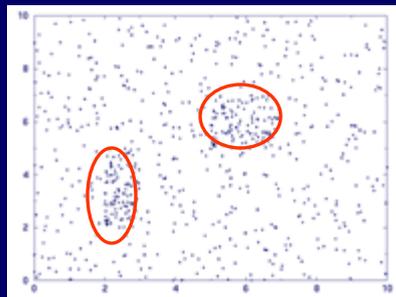


# Data Clustering: an example

- Regulatory binding-sites are short **conserved** sequence fragments in promoter regions

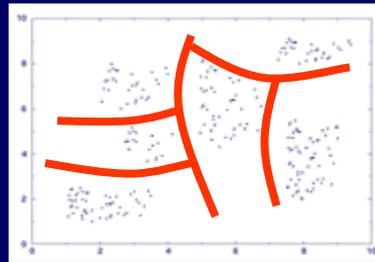
```
..... acgtttataatggcg .....  
.....ggctttatattcgtc .....  
.....ccgatataatcta .....
```

- Solving binding-site identification as a clustering problem
  - ↳ Project all fragments into Euclidean space so that similar fragments are projected to nearby positions and dissimilar fragments to far positions
  - ↳ **Observation:** conserved fragments form “clusters” in a noisy background

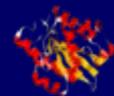
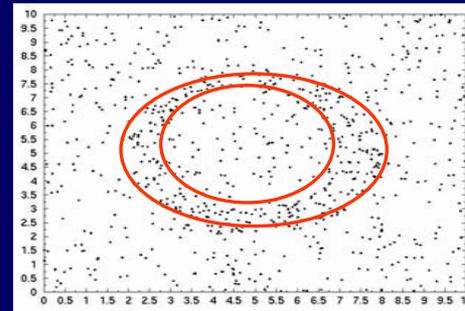
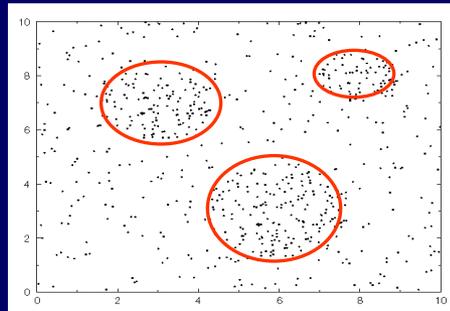


# Data Clustering Problems

- **Clustering:** partition a data set into clusters so that data points of the same cluster are “similar” and points of different clusters are “dissimilar”

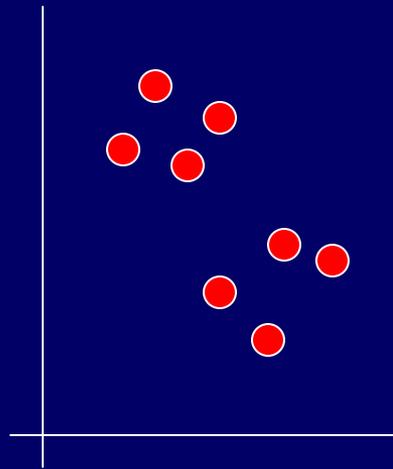


- **cluster identification** -- identifying clusters with significantly different features than the background

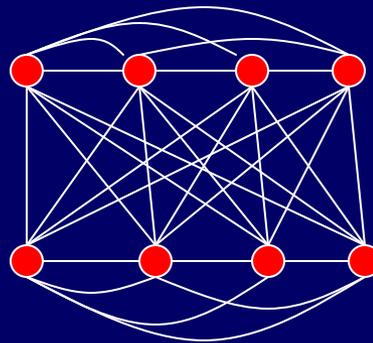


# A Theoretical Framework

- Representation of a set of n-dimensional (n-D) points as a graph
  - ↳ each data point represented as a **node**
  - ↳ each pair of points represented as an **edge** with a weight defined by the “distance” between the two points



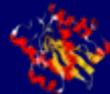
n-D data points



graph  
representation

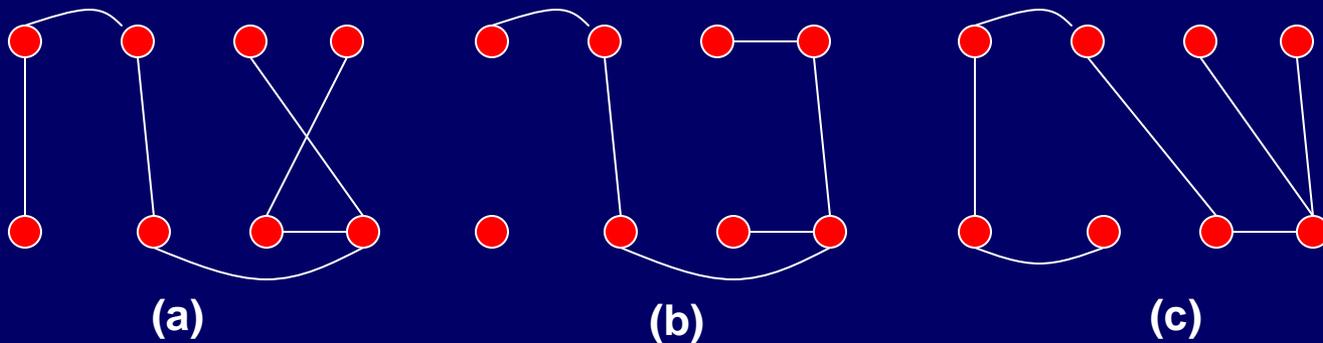
0	1	1.5	2	5	6	7	9
1	0	2	1	6.5	6	8	8
1.5	2	0	1	4	4	6	5.5
				.			
				.			
				.			

distance matrix

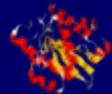


# A Theoretical Framework

- **Spanning tree**: a sub-graph that has all nodes connected and has no cycles

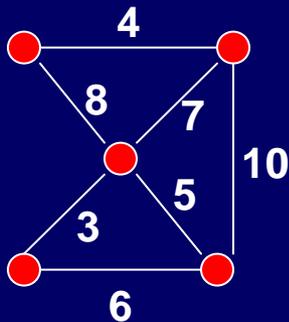


- **Minimum spanning tree**: a spanning tree with the minimum total distance



# A Theoretic Framework

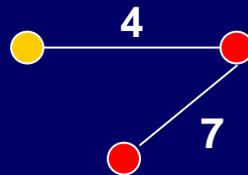
- Prim's algorithm (graph, tree)
  - ↳ step 1: select an arbitrary node as the current tree
  - ↳ step 2: find an external node that is closest to the tree, and add it with its corresponding edge into tree
  - ↳ step 3: continue steps 1 and 2 till all nodes are connected in tree.



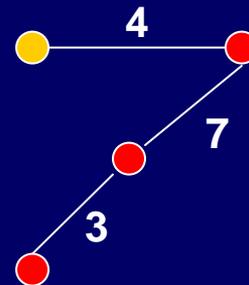
(a)



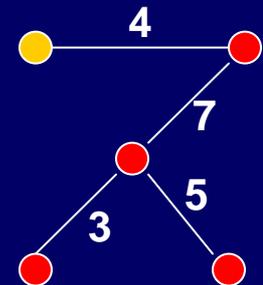
(b)



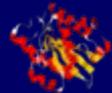
(c)



(d)

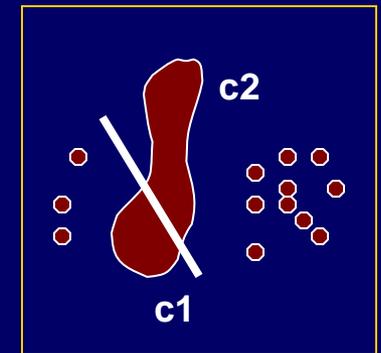
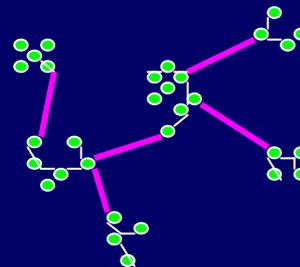
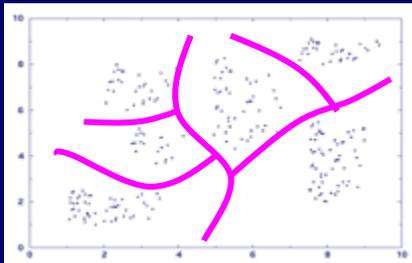


(e)



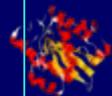
# A Theoretical Framework

- A formal definition of a cluster:
  - ↳  $C$  forms a cluster in  $D$  **only if** for any partition  $C = C1 \cup C2$ , the closest point, from  $D - C1$ , to  $C1$  is from  $C2$ .



- Key results

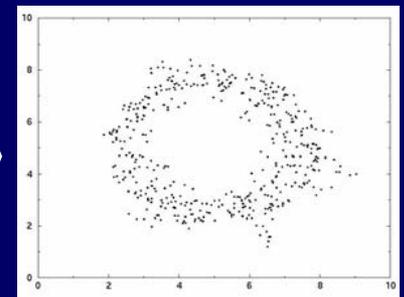
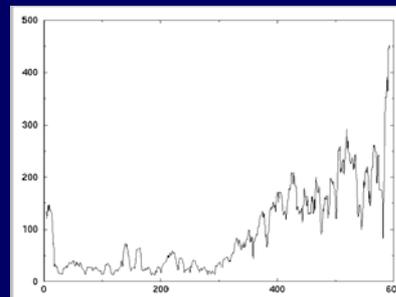
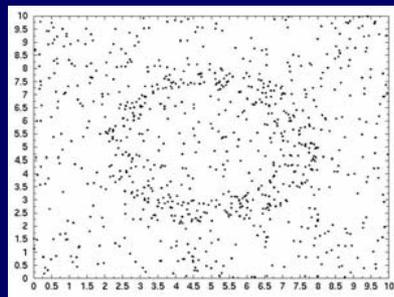
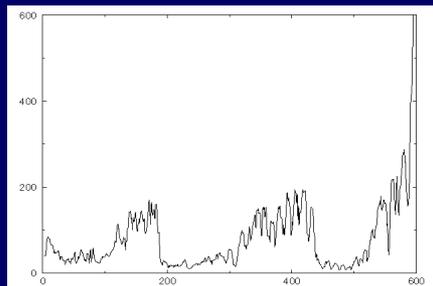
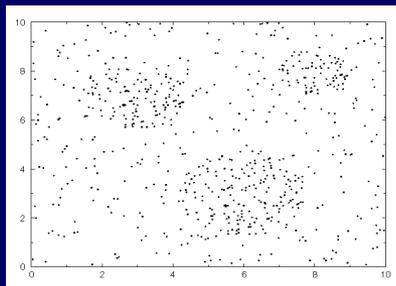
For any data set  $D$ , any of its cluster is represented by a sub-tree of its MST



# A Theoretical Framework

- The selection order of nodes by PRIM's algorithm defines a linear representation,  $L(D)$ , of a data set  $D$

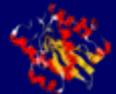
Any contiguous block in  $L(D)$  represents a cluster if and only if its elements form a sub-tree of the MST, plus some minor additional conditions (each cluster forms a valley)



# A Theoretical Framework

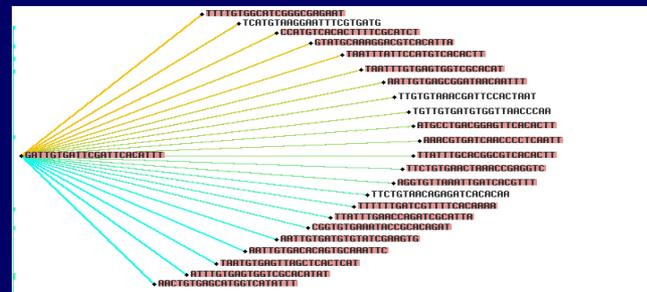
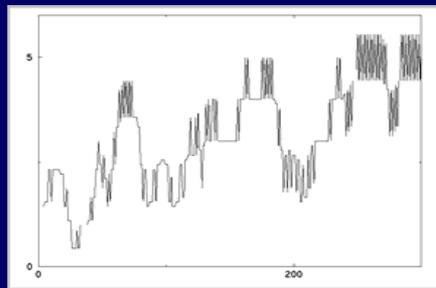
---

**Many biological data analysis problems can be rigorously and reliably solved as sub-string search problems, which we know how to solve!!!**

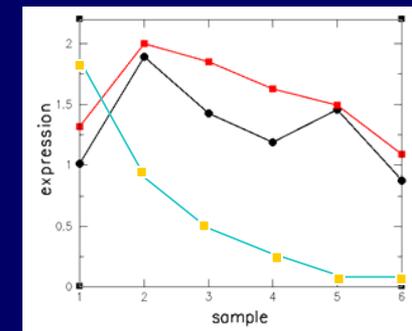
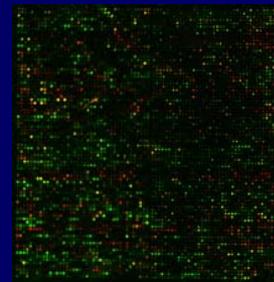
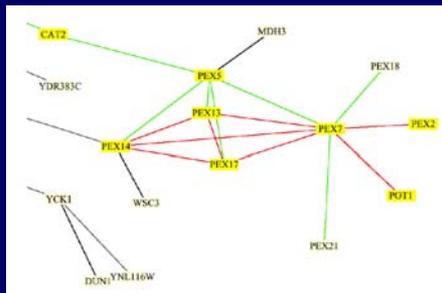


# Application Examples

- Regulatory binding site identification: CRP binding site



- Two hybrid data analysis
- Gene expression data analysis

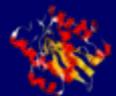


**Are all solvable by the same algorithm!**

# More Application Examples

---

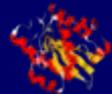
- Phylogenetic tree clustering analysis
- Protein sidechain packing prediction
- Assessment of prediction reliability of protein structures
- Protein secondary structures
- NMR peak assignments
- .....



# What we have learned .....

---

- General common solution may exist for many seemingly unrelated biological analysis problems
- Need more basic research into the data analysis and data mining problem
- Developing these general analysis tools can save time/pain for individual (GTL) researchers from finding/developing tools for their applications



# Infrastructure for Data Analysis and Data Mining

- Identify a set of fundamental problems that cover many important biological data analysis and mining problems
- Implement these fundamental algorithms as a set of (platform-independent) library functions like LINPACK for linear algebra
- Execution of these “library functions” on DOE supercomputers so individual (GTL) researchers can call them as subroutines through internet

**A DOE Data Analysis Center in support of GTL?**

