DATA MINING IN BIOINFORMATICS

DHIFLI WAJDI

POSTDOCTORAL RESEARCHER AT
UNIVERSITY OF QUEBEC AT MONTREAL (UQAM)

DHIFLI.WAJDI@COURRIER.UQAM.CA
HTTPS://SITES.GOOGLE.COM/SITE/WAJDIDHIFLI/

THE NEED FOR DATA MINING IN BIOINFORMATICS

BGI Hong Kong, Tai Po Industrial Estate, Hong Kong

High-throughput technologies:

- Genome and RNA sequencing
- Compound screening
- Genotyping chips
- Bioimaging

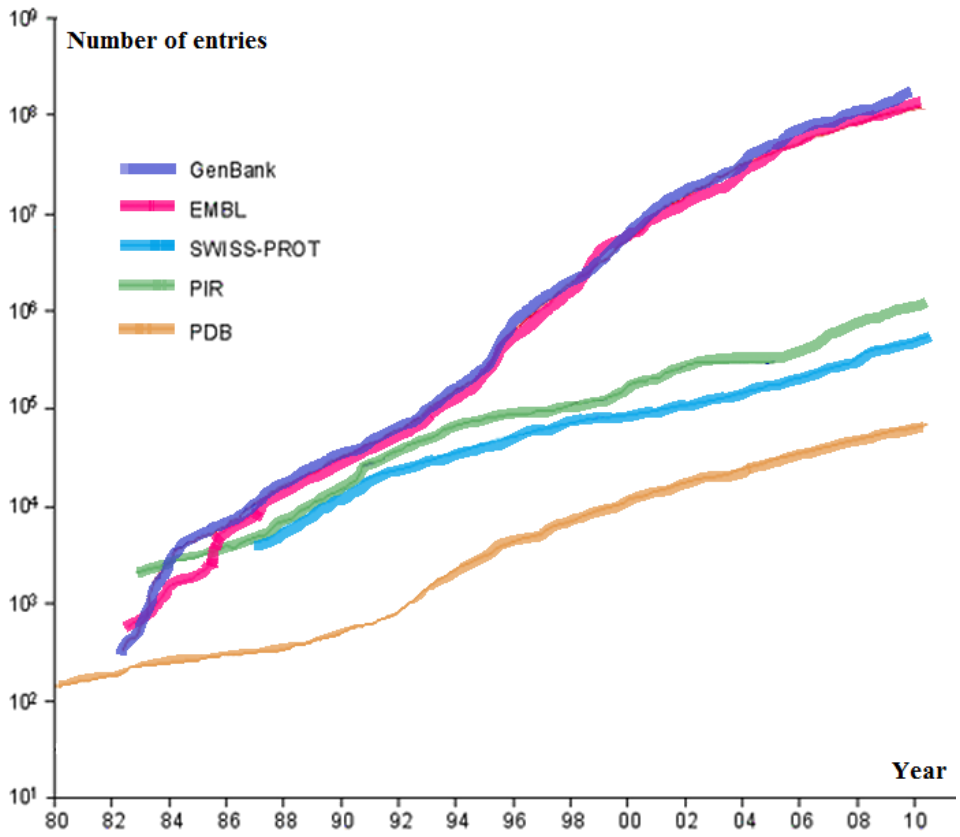
Molecular databases are growing much faster than our knowledge of biological processes.

THE NEED FOR DATA MINING IN BIOINFORMATICS

- **Large collections of molecular data**
  - Gene and protein sequences
  - Genome sequence
  - Protein structures
  - Chemical compounds

- **Problems in Bioinformatics**
  - Predict the function of a gene given its sequence
  - Predict the structure of a protein given its sequence
  - Predict the boundaries of a gene given a genome segment
  - Predict the function of a chemical compound given its molecular structure
  - ….

THE NEED FOR DATA MINING IN BIOINFORMATICS



- Manual lab works are **no longer able to match** the increasing load of data
- The need of **automated, fast and accurate computational** tools is all the more **urgent**

THE NEED FOR DATA MINING IN BIOINFORMATICS

- **Additional challenges**
  - Highly complex
  - Noisy
  - Inconsistent
  - Redundant
  - ….

Data Mining can Help !

DATA MINING

What is data mining?

[Fayyad 1996]: *"Data mining is the application of specific algorithms for extracting patterns from data".*

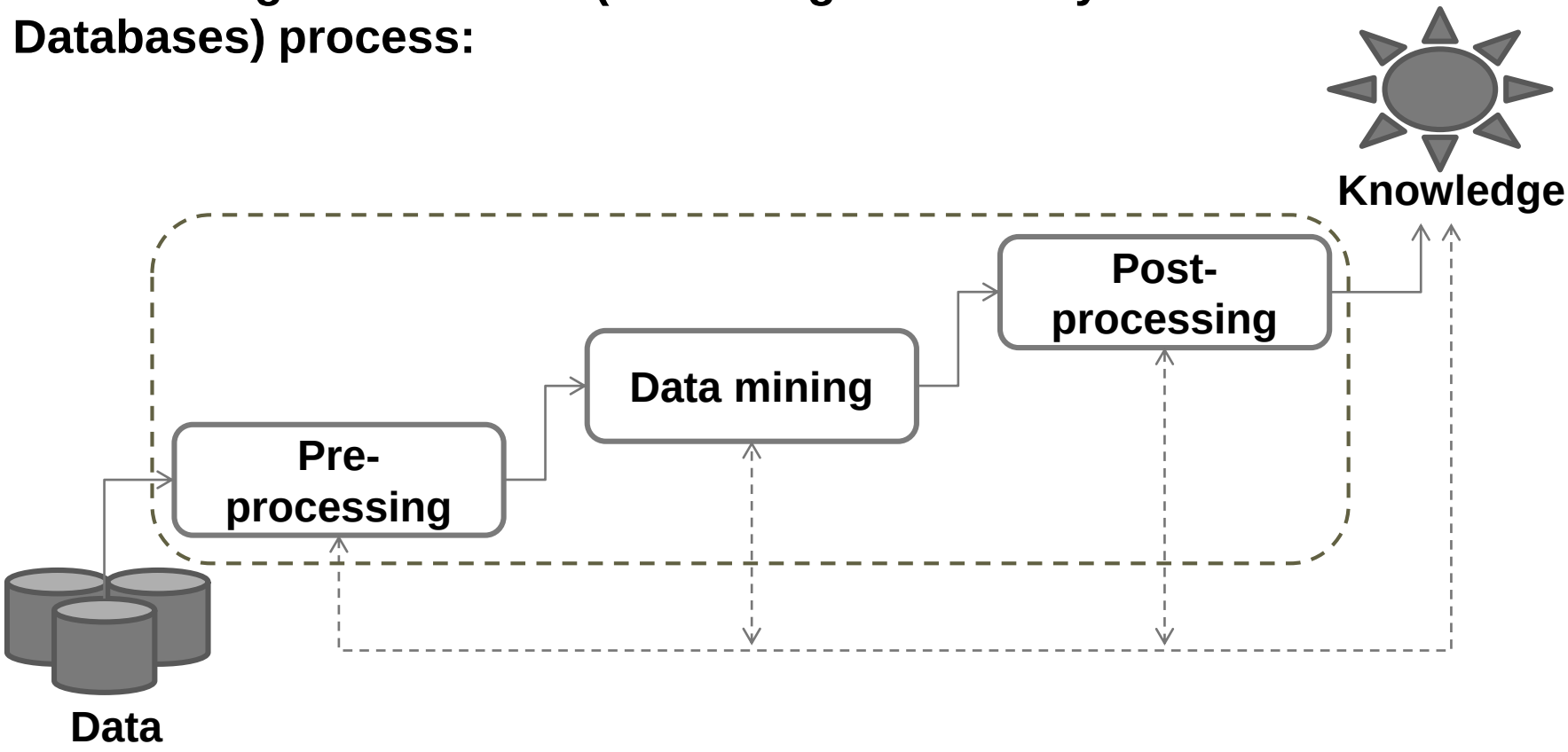
[Han&Kamber 2006]: *"data mining refers to extracting or mining knowledge from large amounts of data".*

[Zaki and Meira 2014]: *"Data mining comprises the core algorithms that enable one to gain fundamental insights and knowledge from massive data".*

Wikipedia: "*it is defined as the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems".*

DATA MINING

Data mining and the KDD (Knowledge Discovery in Databases) process:



DATA MINING

Pre-processing:

- data cleaning (to remove noise)
- data integration (combine multiple data source)
- data selection (to extract subsets of data that are concerned by the analysis)
- data transformation (to transform data into convenient formats that are required from data mining algorithms).

Data mining:

- applying computational methods to extract knowledge from data.

Post-processing:

- evaluation
- validation
- interpretation of the discovered knowledge

DATA MINING

What do we concretely do?

- Clustering
  - grouping of similar objects into sets (K-means, mean-shift, DBSCAN, Spectral Clustering, …)
- Classification
  - Identifying to which category an object belongs to (SVM, KNN, Decision Trees, Neural Networks, …)
- Pattern mining
  - Finding existing (hidden) patterns in data (associations, correlations, subsequences, subgraphs, ….)
- And more …

DATA MINING

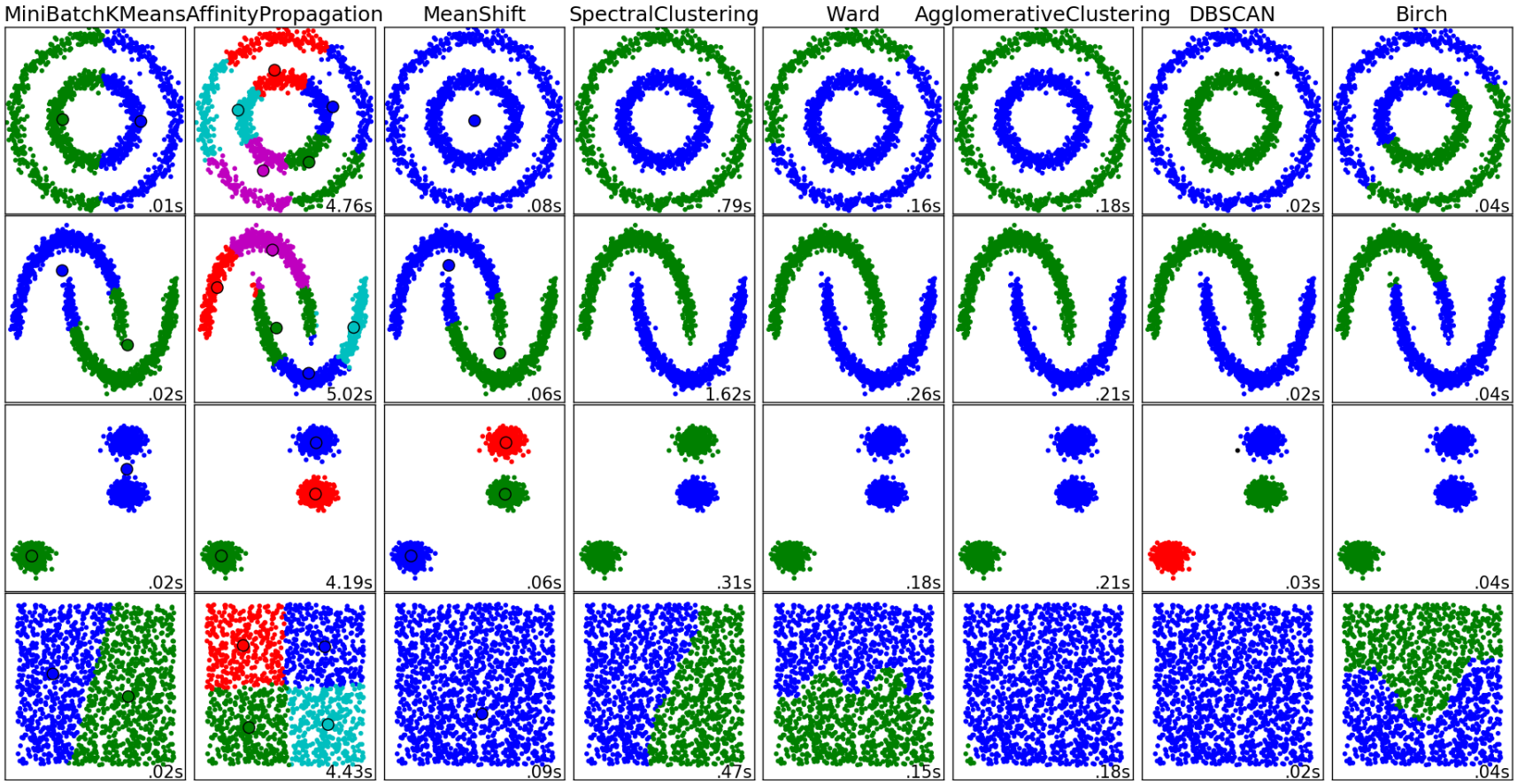
Clustering: K-means

- Partitional clustering approach
- Each cluster is associated with a centroid (center point)
- Each point is assigned to the cluster with the closest centroid
- Number of clusters, K, must be specified
- The basic algorithm is very simple

```
Select K points as the initial centroids.
repeat
    Form K clusters by assigning all points to the closest centroid.
    Recompute the centroid of each cluster.
until The centroids don't change
```

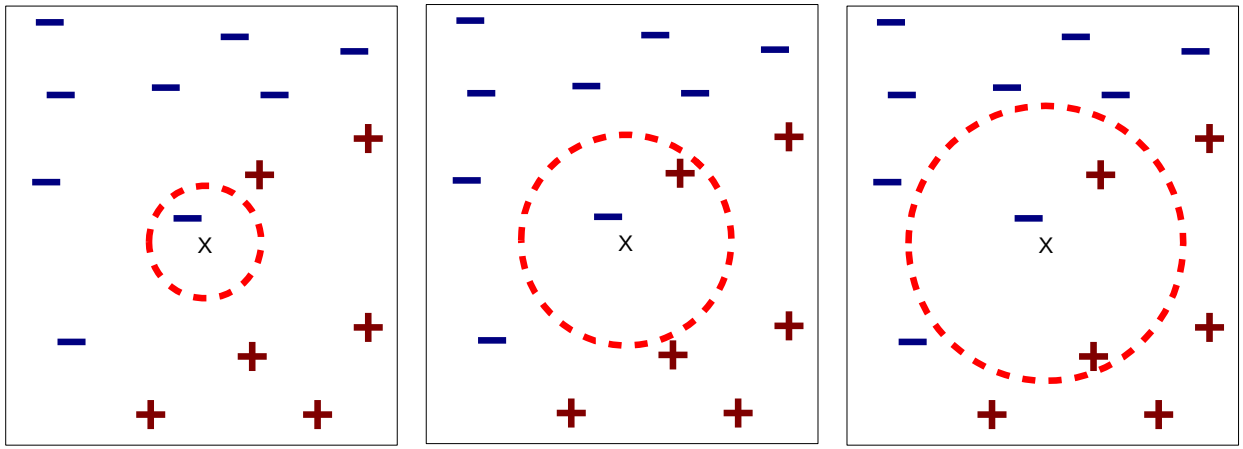
DATA MINING

Remark: Different algorithms can give different Clustering!



DATA MINING

Classification: K-Nearest Neighbors

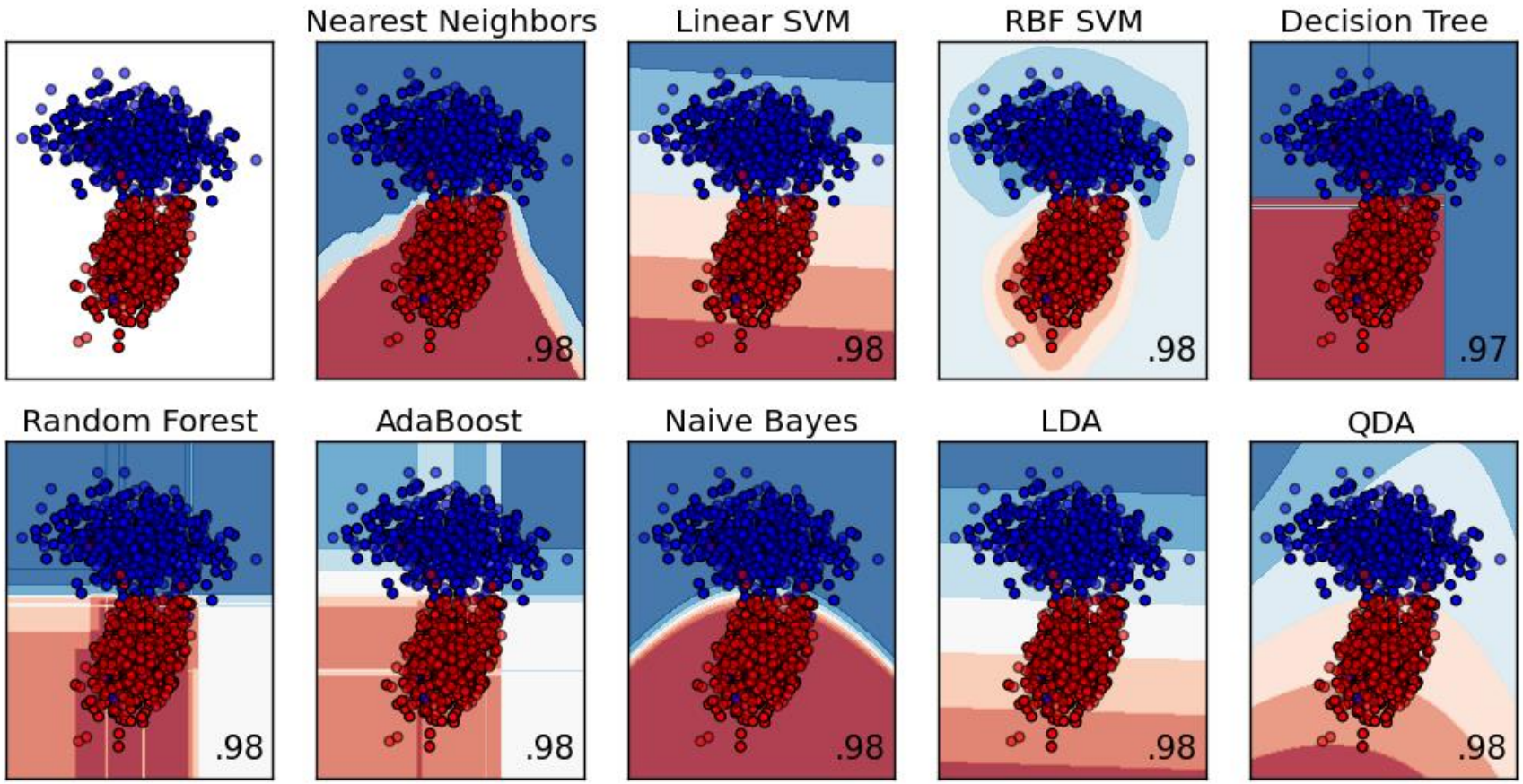


(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

1. Compute distance between two points:
2. K-nearest neighbors of a record x are data points that have the k smallest distance to x
3. Takes the majority vote of class labels among the k-nearest neighbors

DATA MINING

Remark: Different algorithms can give different Classification!

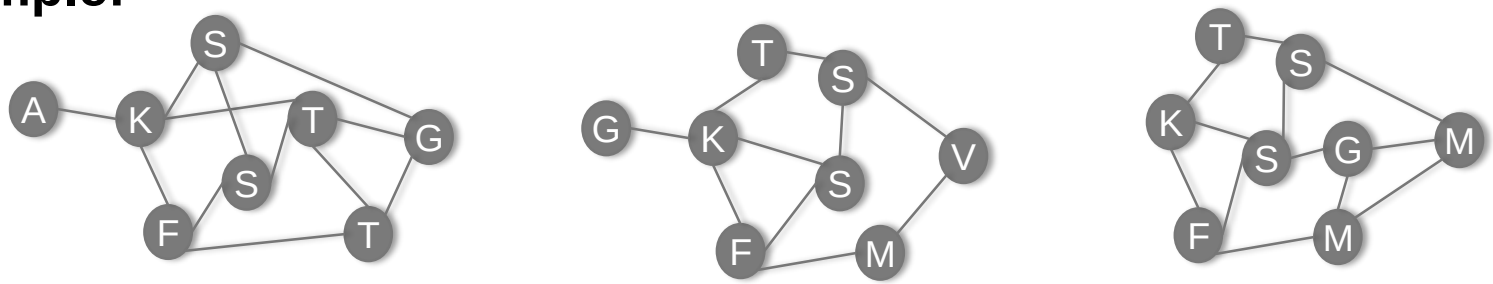


DATA MINING

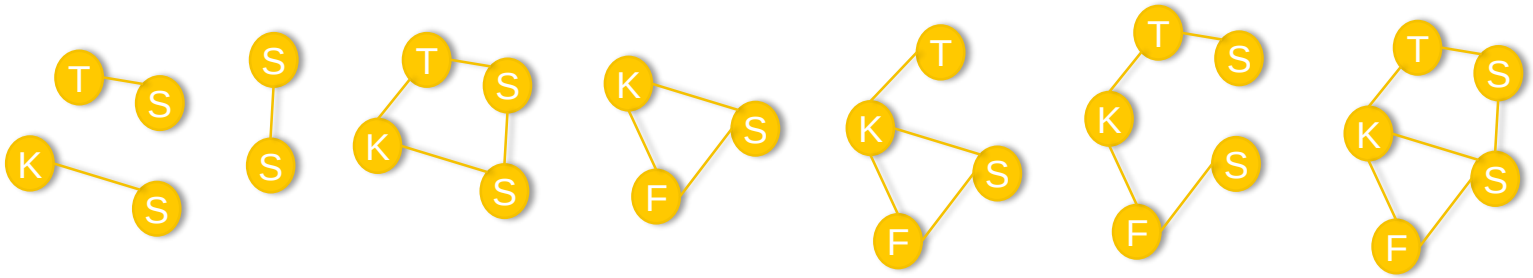
Pattern Mining: Frequent Subgraph Mining

Finding subgraphs that occur in graph data, giving a minimum support

Example:



Graph database



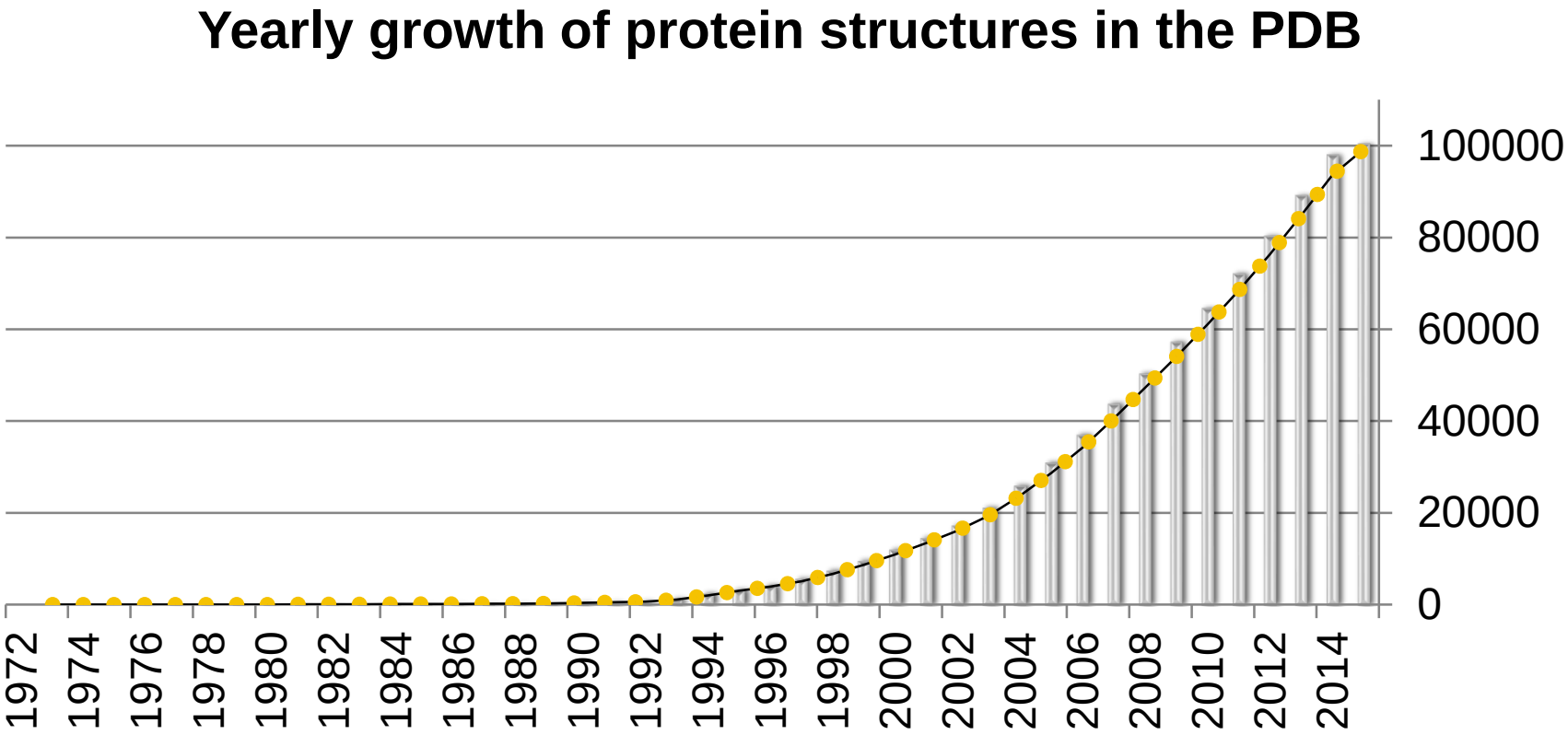
Frequent Subgraphs
(minimum support = 3)

DATA MINING IN BIOINFORMATICS: EXAMPLE 1

PGR

PROTEIN GRAPH REPOSITORY

PGR: PROTEIN GRAPH REPOSITORY

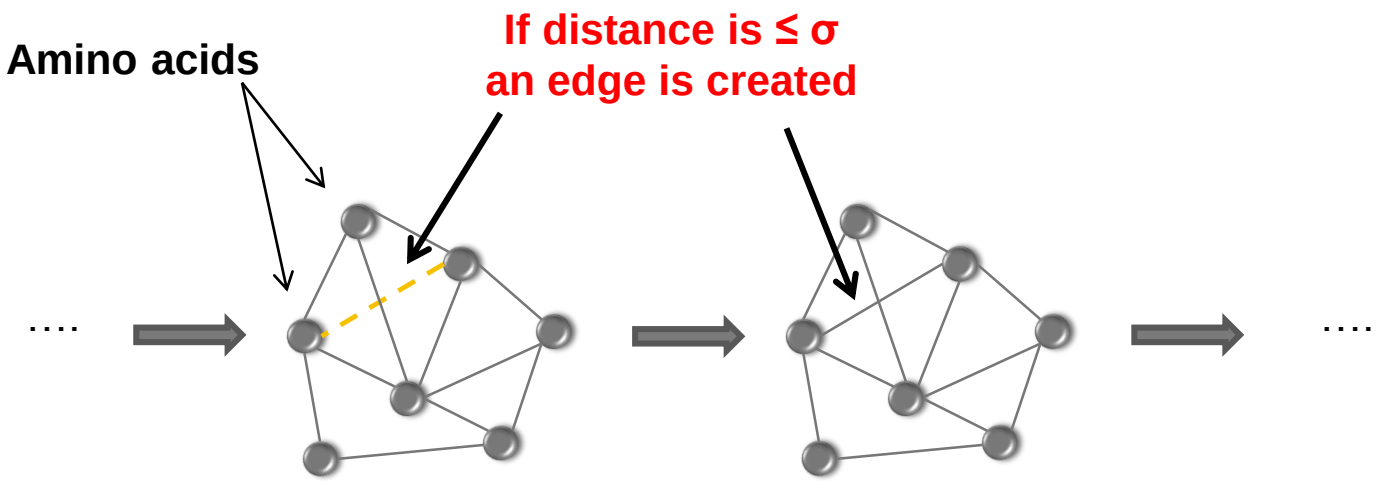


Need of automatic tools to meet the increasing load of data !!

PGR: PROTEIN GRAPH REPOSITORY

A protein 3D structure can be represented by a graph (protein contact map)

- Amino acids ➔ Nodes (labeled with the amino acid type)
- Connections between amino acids ➔ Edges

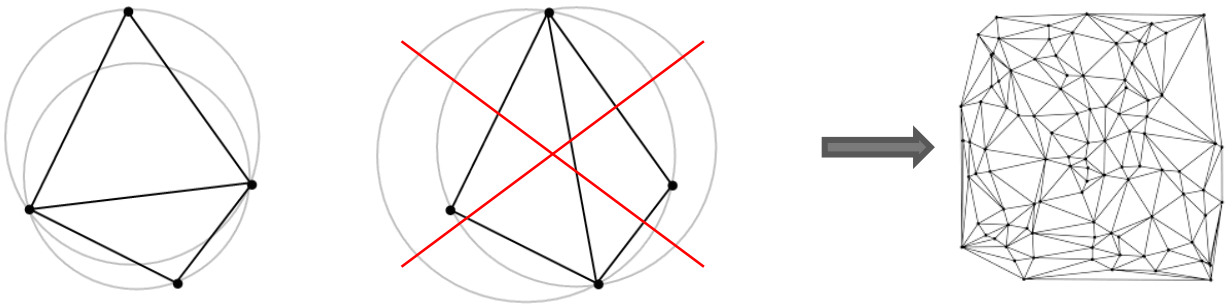


➔ **Use graph mining techniques for automated analysis of protein 3D structure**

PGR: PROTEIN GRAPH REPOSITORY

Protein-to-graph transformation techniques:

- **Delaunay triangulation**

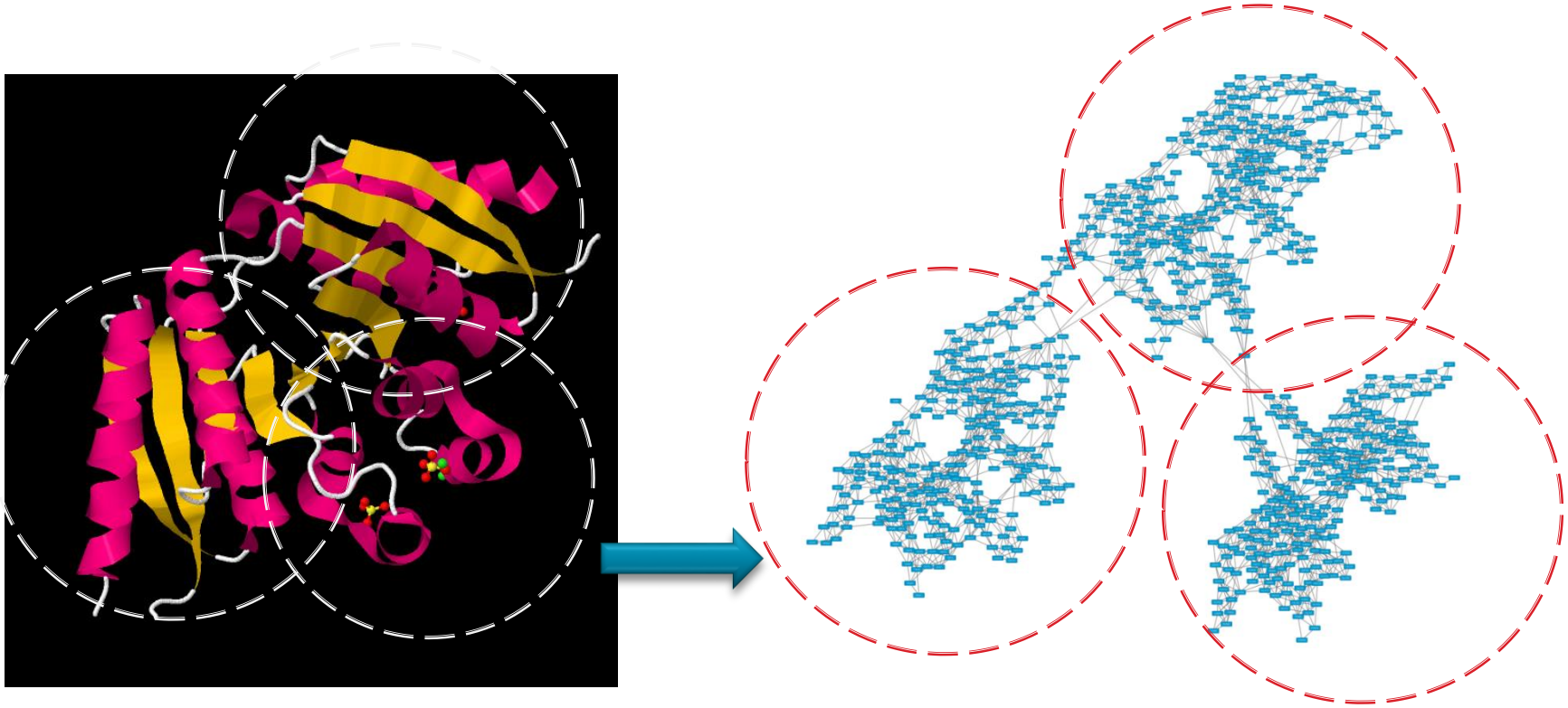


- **Main Atom**
  - Abstract each residue in one atom of it; usually C-alpha atom
  - Two residues are linked by an edge if the distance between their main atoms <= threshold
- **All Atoms**
  - Two residues are linked by an edge if the distance between any pair of their atoms <= threshold

PGR: PROTEIN GRAPH REPOSITORY

A real world example:

A protein 3D-structure (PDB-id: 5AHW) and its corresponding graph (Main Atom, C-alpha).



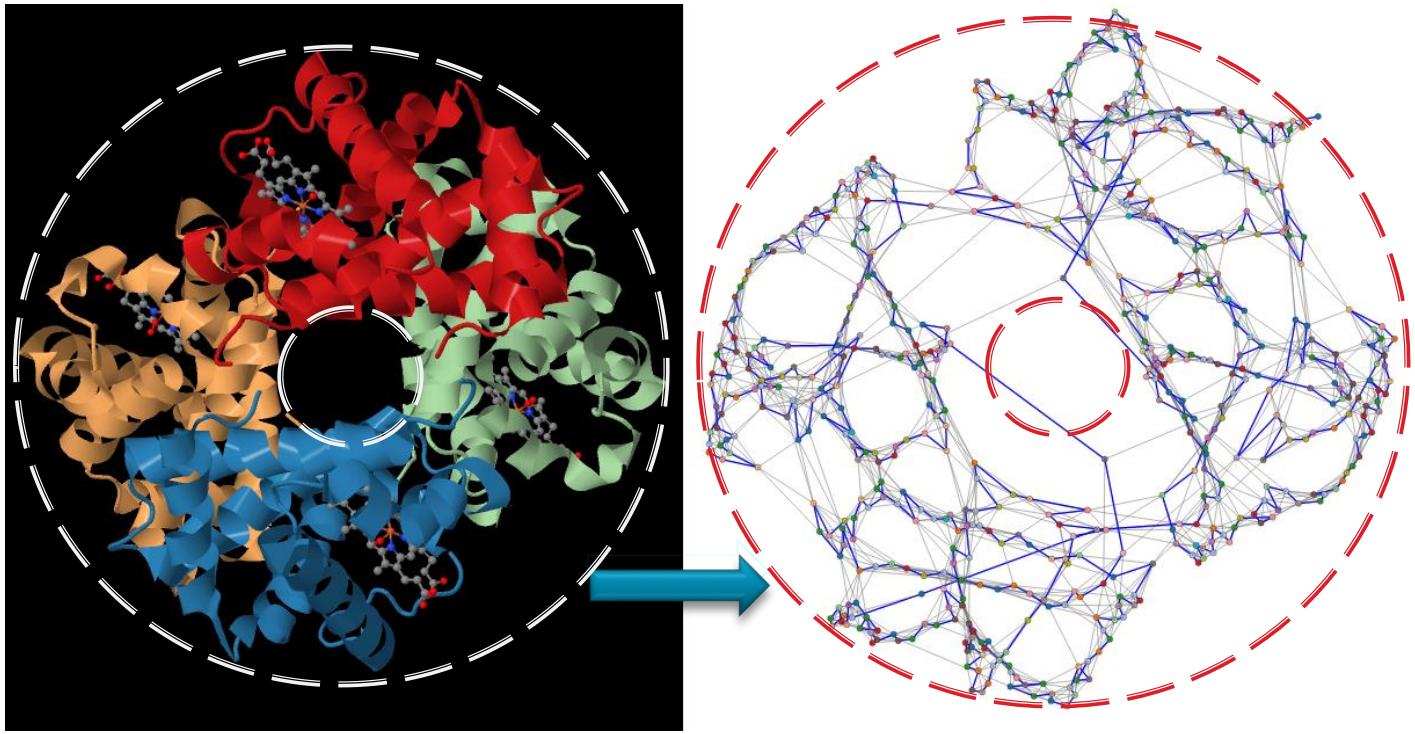
Protein 3D structure

Protein graph

PGR: PROTEIN GRAPH REPOSITORY

A real world example:

The *Human Hemoglobin* protein 3D-structure (PDB-id: 1GZX) and its corresponding graph (Main Atom, C-alpha).



Protein 3D structure **Protein graph**

PGR: PROTEIN GRAPH REPOSITORY

Protein Graph Repository

PG-converter

- Transform protein 3D structure into graph
- Several transformation methods
- Several output formats
- Download and visualization

Repository

- Download protein graphs
- Visualization
- Pre-computed Graph attributes

PG-Similarity

- Find protein structural neighbors
- Distribution of topological attributes for the query protein and its structural neighbors

PGR: PROTEIN GRAPH REPOSITORY

PGR v1.0: Latest Release Statistics Based on PDB dated July 11, 2014	
Number of currently holding proteins graphs	188 252
Number of unique proteins 3D-structures	94 126
Protein graphs based on Cα	94 126
Protein graphs based on All Atoms	94 126

PGR: PROTEIN GRAPH REPOSITORY

Online DEMO:

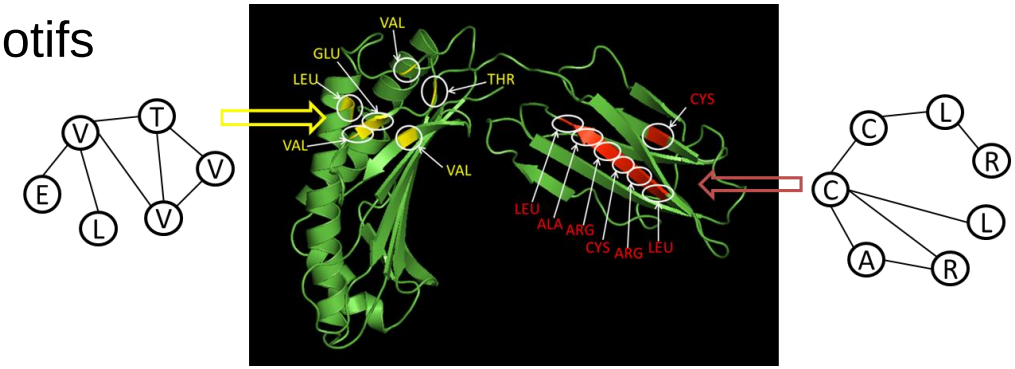
http://wjdi.bioinfo.uqam.ca/

PGR: PROTEIN GRAPH REPOSITORY

Usefulness of PGR

- **Structure alignment and comparison:**
  - Graph Matching: maximal Clique (Vast / Vast+), Edit distance, Similarity measures (MCS)
  - Graph Embedding: Topological similarity

- **Pattern mining:**
  - Subgraphs (frequent, discriminative, …)
  - Fingerprints
  - Functional / structural motifs
  - Binding sites
  - ….



DATA MINING IN BIOINFORMATICS: EXAMPLE 2

ProtNN

Fast and Accurate Protein 3D-structure classification in Structural and Topological Space

PROTNN: FAST PROTEIN 3D-STRUCTURE CLASSIFICATION

Existing protein Classification techniques:

- **Sequence-based classification (Blast, ProtFun, SVM-Prot, ...)**
- **Structure-based classification (*e.g.* Combinatorial Extension, Sheba, FatCat, Fragbag, ...)**
- **Subsequences / substructures-based classification**
  - The subsequences / substructures are used as features to identify the function of unknown proteins

PROTNN: FAST PROTEIN 3D-STRUCTURE CLASSIFICATION

- Sequence (and subsequences)-based classification **do not** incorporate spatial information !
  - less efficient in classifying structurally similar proteins with low sequence similarity

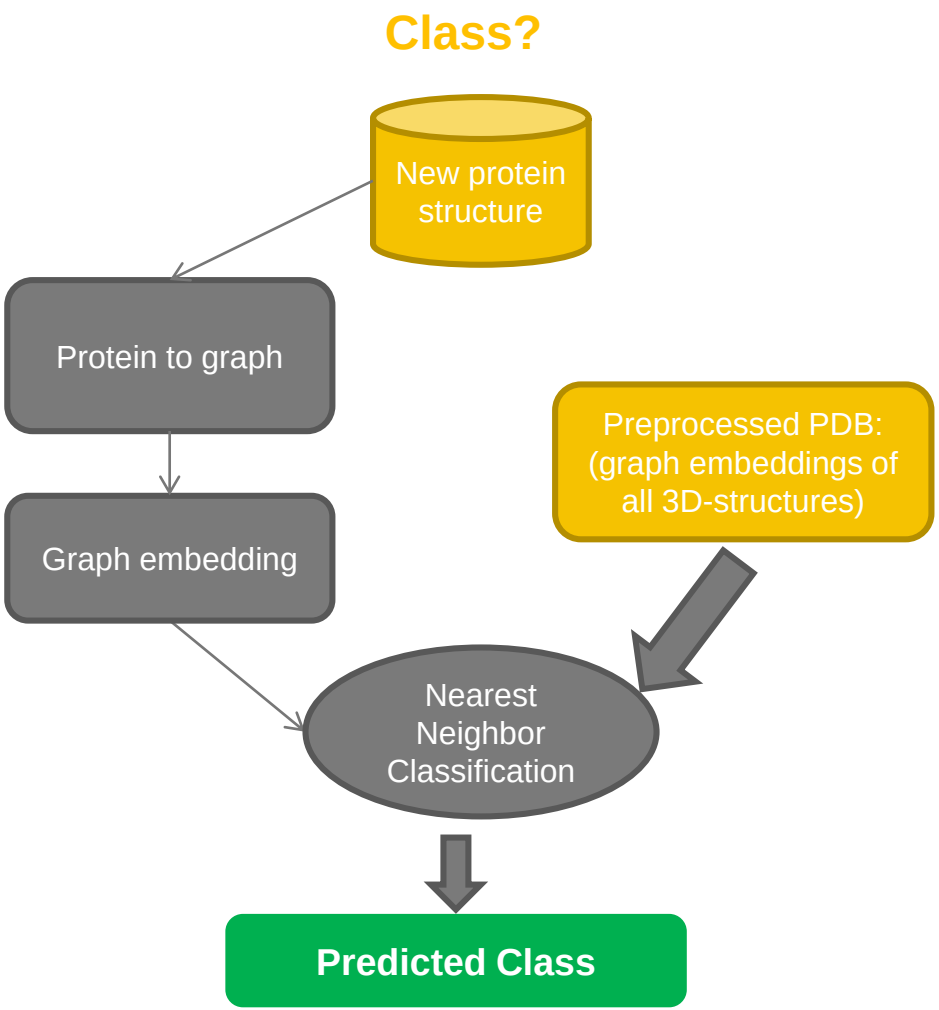
- Structure and substructure-based classification techniques **do** incorporate spatial information
  - But suffer computational cost!

⇨ It is essential to find an **efficient** way to **incorporate 3D-structure information** with **low computational complexity**

PROTNN: FAST PROTEIN 3D-STRUCTURE CLASSIFICATION

General Framework

- We used a set of 18 structural, topological, spectral, and label graph attributes



PROTNN: FAST PROTEIN 3D-STRUCTURE CLASSIFICATION

Results

Accuracy comparison of ProtNN with other classification techniques.

Dataset	Classification approach								
	Blast	**Sheba**	**FatCat**	**CE**	**LPGBCMP**	**D&D**	**GAIA**	**ProtNN**	**ProtNN***
DS1	0.88	0.81	**1**	0.45	0.88	0.93	**1**	0.97	0.97
DS2	0.82	0.86	**0.89**	0.49	0.73	0.76	0.66	0.8	**0.89**
DS3	0.9	0.95	0.84	0.59	0.90	0.96	0.89	0.96	**0.97**
DS4	0.76	0.92	**1**	0.46	0.9	0.93	0.89	0.97	0.97
DS5	0.86	**0.99**	0.94	0.76	0.87	0.89	0.72	0.9	0.94
DS6	0.78	**1**	0.94	0.81	0.91	0.95	0.87	0.96	0.96
Avg. accuracy[1]	0.83±0.05	0.92±0.07	0.94±0.06	0.59±0.15	0.86±0.06	0.9±0.07	0.84±0.12	0.93±0.06	**0.95±0.03**
Avg. distances[2]	0.14±0.07	0.05±0.07	0.04±0.05	0.38±0.15	0.11±0.03	0.7±0.04	0.14±0.09	0.05±0.03	**0.02±0.01**
Rank	8	4	2	9	6	5	7	3	**1**

[1] Average classification accuracy of each classification approach over the six datasets.
[2] Average of the distances between the accuracy of each approach and the best obtained accuracy with each dataset.

PROTNN: FAST PROTEIN 3D-STRUCTURE CLASSIFICATION

Results

Runtime results of ProtNN, FatCat and CE on the entire Protein Data Bank.

Task	Total runtime[1]	Runtime[1]/protein
Building graph models	23h:9m:57s	0.9s
Computation of attributes	5d:8h:12m:29s	4.9s
Classification	2h:55m:15s	0.1s
ProtNN (all)	6d:10h:17m:41s	5.9s
FatCat	Forever[2]	2d:4h:24m:19s[3]
CE	Forever[2]	2d:10h:7m:2s[3]

[1]The runtime is expressed in terms of days:hours:minutes:seconds
[2]The program did not finish running within two weeks
[3]The average runtime of randomly selected 100 proteins

THANKS

QUESTIONS