UCS2735 BIOINFORMATICS TECHNOLOGIES

Unit 2

Datawarehousing and Datamining in Bioinformatics

1. Data Warehouse Architecture
2. XOD
3. Data Mining in Bioinformatics
4. Symbols in Sequence Alignment
5. Gene Prediction using similarity-based Methods
6. Phylogenetic Tree Construction
7. Protein Prediction Methods
8. GOR Method
9. Nearest Neighbor Method
10. Neural Network Protein Prediction
11. Threading for 3D structure of proteins

* Data Warehouse for Bioinformatics

→ A biological data warehouse is a subject-oriented, integrated, non-volatile, expert interpreted collection of data in support of biological data analysis and knowledge discovery.

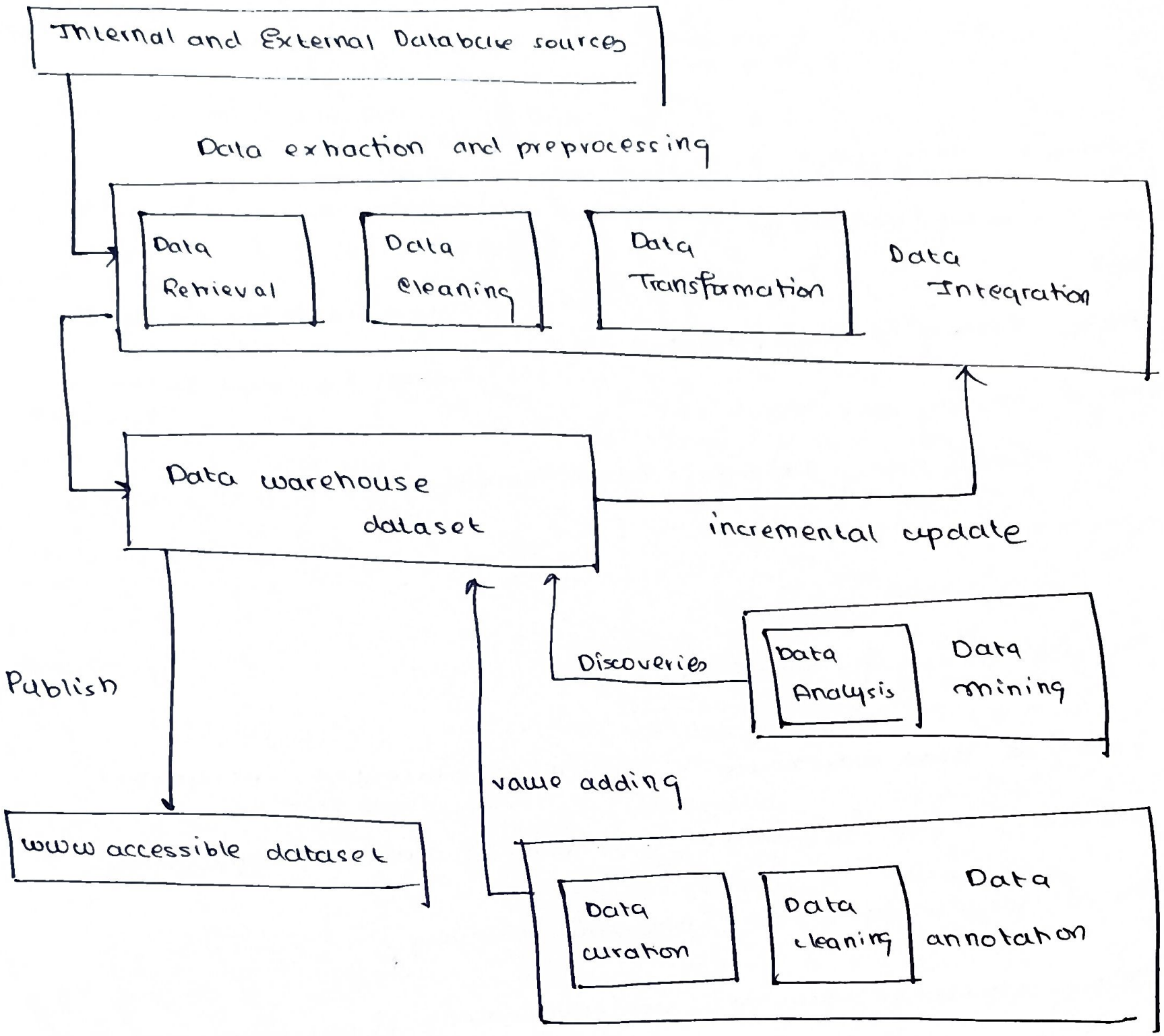
→ A data warehouse provides an environment of reusable data integration mechanisms and data ~~cleaning~~ cleaning support to manage and organize assorted sets of specialized data for different data mining purposes

→ Steps involved in data warehousing processes are data integration, data cleaning, and data analysis.

→ can mention about DSS

Data Warehouse Components : (i) Retrieval of data from databases
(ii) mechanism of cleaning data
(iii) flexibility of manipulating datasets
(iv) integrating and designing purposeful analysis tools that can be used jointly or independently

Architecture



Data Mining Tools & DBs

DNA Sequence Databases – GenBank (NCBI), EMBL (European Molecular Biology Laboratory)

Protein Sequence DB – Swiss-Prot, PIR, TrEMBL

Entrez – NCBI – identify mRNA, literature, protein records, conserved domains, mutations

Sequence Comparison & Alignment – FASTA, BLAST, Clustalw

Vizualization

* Transforming Data to Knowledge

→ Transformation of data to knowledge, is also known as knowledge discovery from databases (KDD).

→ KDD is the non-trivial extraction of implicit, previously unknown and potential useful information from data.

* Data Mining in Bioinformatics

→ The software tools that facilitate research in bioinformatics are broadly categorized into four:

① Data Retrieval Tools → Entrez
→ integrated system by NCBI
→ has literature, nucleotide and protein sequences, complete genomes, 3D structures etc.

② Sequence Comparison & Alignment Tools → BLAST
→ FASTA
→ Multiple sequence alignment → ClustalW

③ Pattern Discovery Tools → used to search for patterns or features in the data
→ an important tool is cluster analysis
→ some integrated pattern discovery tools are Expression Profiler and Gene Quiz

④ Visualization Tools → interactive, graphical display of genomic data
→ have visualization integrated within them
→ Available visualization tools are Tree View, BioViews

* DNA Data Analysis

4 bases : Adenine (A)

Cytosine (C)

Guanine (G)

Thymine (T)

Complementary pairings

A ↔ T

G ↔ C

DNA $\underset{\text{reverse transcription}}{\overset{\text{transcription}}{\rightleftarrows}}$ RNA ⟶ Protein

→ DNA data analysis includes sequence comparison

&

sequence alignment

- pairwise alignment (compare 2)
- multiple sequence alignment

In global sequence alignment - align the entire sequence to try to maximize the degree of similarity

In local sequence alignment, the alignment stops at the ends of regions of strong similarity, and a much higher priority is given to finding these local regions than to extending the alignment to include more neighboring pairs.

Methods → Dynamic programming

→ Needleman-Wunsch

→ Smith-Waterman

* Symbols in Sequence Alignment

Gaps : — ⇒ insertions and deletions in one sequence compared to the other

Asterisk : * → indicates identical amino acid residues - perfect matches across all sequences at that position

Colon : : ⇒ Represents a conserved substitution where the amino acids have similar properties like polarity or size

Dot : . ⇒ Indicates a semi-conserved substitution, meaning the amino acids at this position have weakly similar properties

* Gene Prediction using similarity-based methods - identify matching subsequences with the given gene and the DNA segment write position

* Phylogenetic Tree → A phylogenetic analysis of a family of related DNA or protein sequences is a determination of how the family might have been derived during molecular evolution

→ Phylogenetic analysis leads to the construction of an evolution tree.

→ Common methods used are: (i) maximum parsimony method
(ii) distance method
(iii) maximum likelihood method

* Phylogenetic Tree Construction — UMGMA method

↓

Unweighted Pair Group Method using the Arithmetic Mean

Consider the following gene sequences:

A - A T C G A T C G

B - G T A G A C G A

C - A C C G T A C G

D - T C A G T C A G

E - G C C T A C A G

	A	B	C	D	E
A		5	3	6	5
B			7	5	6
C				4	5
D					4
E					

A -	A	T	C	G	A	T	C	G
B -	G	T	A	G	A	C	G	A
C -	A	C	C	G	T	A	C	G
D -	T	C	A	G	T	C	A	G
E -	G	C	C	T	A	C	A	G

choose min = 3 → combine A/C

	A/C	B	D	E
A/C		6	5	5
B			5	6
D				4
E				

$(A/C, B) = \frac{|A-B| + |C-B|}{2} = \frac{5+7}{2} = 6$

$(A/C, D) = \frac{|A-D| + |C-D|}{2} = \frac{6+4}{2} = 5$

$(A/C, E) = \frac{|A-E| + |C-E|}{2} = \frac{5+5}{2} = 5$

choose next min = 4

	A/C	B	D	E
A/C		6	5	5
B			5	6
D				1
E				

combine D/E

	A/C	D/E	B
A/C		5	6
D/E			5 5
B			

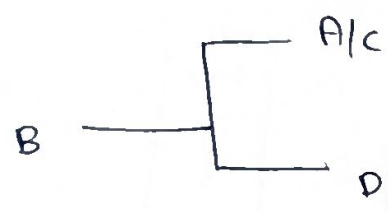
$$(A/C, D/E) = \frac{|A/C - D| + |A/C - E|}{2}$$

$$= \frac{5+5}{2} = 5$$

$$(D/E, B) = \frac{|B - D| + B \cdot E|}{2}$$

$$\frac{5+6}{2} = 5.5$$

min = 5.5 (look at unaveraged vals)



B closer to D/E

see CAT 1 answer key

* Protein Analysis and Prediction

Methods used →

(1) Comparative Modelling → relies on the similarity between sequences of unknown and known proteins to predict structure and prediction

→ use tools and databases like PFAM and PROSITE

(2) Ab Initio Prediction → predicts protein structures directly from the amino acid sequence using methods like:

A. Statistical-based Feature Methods — GOR method

- Garnier Osguthorpe and Robson Method

- sequence is scanned using a sliding window for the occurrence of amino acids that have a high probability - and determining likelihood

B. Nearest-neighbor Methods - identify training sequences of known structures that are homogenous to the query sequence

C. Neural Networks - use sliding windows to train and predict structures based on sequence patterns - eg. PHDsec

(3) HMM models

* GOR Method for Protein Prediction

→ There will be a scoring matrix and corresponding probabilities given.

→ For the bases in the subsequence - find the sum vertically

$\leq P(H)$, $\leq P(E)$, $\leq P(C)$

→ Find info gain value - max(other two values)
→ Choose the alphabet with highest value
} equivalent to finding max

→ Slide the window, repeat

→ Choose the secondary structure that occurs the most no. of times

* Nearest Neighbor Method - Protein Prediction

Seq ID	Sequence	Secondary Structure
T_1	ARHTE	CCHHE
T_2	RHTEC	CHHEC
T_3	HTECL	HHECH

Query Sequence : Q_1 HTECR

len = 5

If training data len > query sequence
chunk in to 2 sequences

find the secondary structures for all the training samples

seq ID	sequence ID	central residue Secondary Structure	secondary structure
T_1	ARHTE	H	H
T_2	RHTEC	T	H
T_3	HTECL	E	E

if k = 3 , match 3 neighbors

		secondary structure
ARHTE →	HTE match	H
RHTEC →	HEC match	H
HTECL →	HTEC match	E

H occurs the most no. of times

choose H

see SAT Q

* Neural network based method of protein predictions

1. Encode features - one hot
2. Define neural network — Input layers
   — Hidden layers
   — Output layers
3. Train
4. Compute error - backpropagation, weight updation
5. choose secondary structure with the highest probability (classification)

* Threading to predict 3D structure of Proteins

→ Threading, also known as fold recognization is a method used to predict the 3D structure of a protein based on its amino acid sequence.

Steps

① Input sequence - The target protein structure we want to predict the 3D structure for

② Template Library - Use a DB of known protein structures usually obtained from experimental data - X-ray crystallography NMR spectroscopy

③ Define scoring criteria — structure - structure compatibility
- structural constraints
- energy considerations

(4) Threading - The target sequence is threaded through the template structure This involves placing the sequence onto the template and optimizing the alignment to best fit the target sequence.

(5) Alignment Optimization - shifting, rotating, deforming to achieve the best fit

(6) Scoring and Score - Thread through multiple templates, score alignments and choose the top models

(7) Model Refinement - techniques like energy minimization, molecular dynamics simulation or further structural adjustments