Current solved paper *BIF 401*

(For Finals)

Created by: Cr. SABA ALWI

Q. Define Top down proteomics?

Ans: "Top down proteomics can handle whole proteins".

Top down proteomics measures the intact proteins followed by peptides after fragmentation.

Q. Need for Chau fasman algorithm?

Ans: Chau fasman algorithm is a propensity base method. The method is based on analyses of the relative frequencies of each amino acid in alpha helices, beta sheets, and turns based on known protein structures solved with X-ray crystallography. From these frequencies a set of probability parameter were derived.

Q. Write down an objective base method and clustering approach?

Ans: UPGMA is a clustering approach and least square distance is an objective base method.

Q. Differentiate between finger printing and short gun proteomics?

Ans: Shotgun proteomics digests the entire protein mix 1st followed by peptide analysis & protein database search

• Peptide mass fingerprinting involves protein separation followed by a single protein's peptide analysis.

Q. what are substitutions and indels?

Ans: Indels:

Removal and addition of amino acids in proteins and nucleotides in DNA, RNA by using Gaps named as Indels.

Substitution:

Replacement of a single or a specific pair of amino acid or nucleotide with other on in a sequence is called substitution.

Substitutions	A**C**GA →□A**G**GA

Q. How can psuedonots detected in the prediction of RNA?

Ans: Tertiary or 3' structure of RNA may form pseudoknots to detect the pseudoknots in RNA structure we need "circular plot" which is a graphical approach.

Q. Formation of protein sequence and limitation?

Ans: There are two basic techniques for sequence of a protein,

1. Edman degradation

2. Mass spectrometery

Edman degradation used to sequence the smaller length protein, about 40 – 50 amino acids.

Large protein sequence is carried out by mass spectrometery.

*Limitations ni mili,

Q. Uniprot and ESEMBL or EMBL?

Ans> **Uniprot** is a public data base which is used to find the sequence of protein.

ESEMBLE is genome search engine which is used to search the genome of every recorded species. OR it might be EMBL so,

Input to FASTA search can be in

* **EMBL*,**= European molecular biology laboratory (EMBL)

Q. Mas spectrometer?

Ans> Mass spectrometer is used to measure mass/charge ratio of ionized proteins and peptides. Data output from the MS comprises of m/z ratios and intensities of each molecule that is measured.

Function of Tandem Ms:

Tandem MS helps in measurement of mass to the fragments as well. This process

provides another step in further scoring and ranking and Protein identification thus becomes easier.

Q. Define Orthology?

Ans> it is a homologous gene present in two organisms that encode protein with same function. It is evolved by direct vertical descent.

Q. What does role of 5´ cap and 3´ poly A tail?

Ans: the 5´ end of molecule caped with 7 methyl guanosine tri phosphate and 3´ end poly A tail cap, both caps play major role to translate mRNA to ribosome, also protect mRNA.

Q.: Explain applications of bioinformatics.

There are some major applications of bioinformatics:

In this field bioinformatics help us in gene finding, in assemblage of genes and forming of databases.

2.Proteomics:

It helps us to decoding protein sequences for better understanding about protein structure, protein to protein relationship, post translational changes and in generating databases for their sequence and structure.

3.Evolutionary study:

It can make phylogenetic trees to find evolutionary relationship between species, also show ancestry between them,

4.System biology:

Bioinformatics also helps us in regulatory mechanism in genes and proteins. So that we can better understand about regulators and treat them by evaluate drugs.

Bioinformatics also help us in further many aspects like,

Drug development

Waste cleanup

Gene therapy

Preventative medicine etc.

Q=write types of phylogenetic trees ? no option

Rooted trees:

- Each node with descendants represents the inferred most recent common

ancestor of the descendants.

- The edge lengths in some trees may be interpreted as time estimates.
- Rooted trees can show temporal evolutionary direction.
- Expensive.

Unrooted:

- Only the relatedness of the leaf nodes.
- Do not require the ancestral root to be known or inferred.
- Less expensive.

Uses:

In bioinformatics, such as

- Rooted and unrooted trees can be used to show phylogenetic relationships

between sequences.

Q. define open reading frame?

In molecular genetics, an **open reading frame** (**ORF**) is the part of a **reading frame** that has the potential

to code for a protein or peptide. An **ORF** is a continuous stretch of codons that do not contain a stop

codon (usually UAA, UAG or UGA).

Q=what is FASTA? Why we use it?

➢ FASTA can perform quick comparison of protein and nucleotide sequence

➢ Can also perform genome and proteome similarities search

➢ Its also available online like BLAST.

FASTA can search sequence databases and identify unknown sequences by comparing them to the

known sequence databases. This can help obtain information on the parent organism, function and

evolutionary history

Q.What is BLAST ? NO OPTION

➢ BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical

significanceIn bioinformatics, **BLAST** for **B**asic **L**ocal **A**lignment **S**earch **T**ool is an algorithm for comparing primary biological sequence information, such as the aminoacid sequences of proteins or the nucleotides of DNA sequences. A BLAST search enables a researcher to compare a query sequence with a library or database of sequences, and identify library sequences that resemble the query sequence above a certain threshold.

➢ Different types of BLASTs are available according to the query sequences. For example, following the discovery of a previously unknown gene in the mouse, a scientist will typically perform a BLAST search of the human genome to see if humans carry a similar gene; BLAST will identify sequences in the human genome that resemble the mouse gene based on similarity of sequence.

Q. What is the difference between basic and acidic amino acids?

Ans: the difference is in their side chains, the acidic ones have acidic side chain while basic one have basic side chain at neutral PH.

• Arginine, lysine and histidine are the basic ones.

• Aspartate and glutamate are the acidic ones.

Q. What is scope of bioinformatics?

Ans. Bif primarily deals with digitalized biological information as well as data reported from bioogy experiments. Computational methods, data processing techniques and algorithms are use in adessing the following issues,

Use of computational algorithms and techniques for:

1. Storage,
2. Organization,
3. Analysis, and
4. Representation of biological information

How junctions are formed?

Ans: junctions are formed when two or more double stranded rigions of RNA converging to forma closed structure.

Q. Name of famous protein data base?

Ams: SWISS PROT and UNI PROT

Q. Four objectives of comparing sequence?

Ans: 1. Similarities among sequences

2. Differences b/w sequences

3. Evolutionary history

4. Predict the function of molecules

Q. How buldges are formed?

Ans: in RNA 2´ structure bulges are formed when double stranded rigion cannot form base pairs perfectly.

Q. how biological simulation help community?

Ans: it helps community by predicting disease, progression and outcomes.

Q. what is role of EXPASY?

Ans: it provides access to a Varity of online data bases and tools depending upon our requirements. We can find sequence information by it.

Q. Define PSTs? No option

Ans: peptide sequence tags (PSTs) are actually small amino acid sequence of peptides producing during MS2.

Q. What is meant by "PK" value of an amino acid?

Ans: PK is the value for an amino acid is that PH at which exactly half of amino acids are charged and half are not charged.

Q. how loops are found after the finding of alpha and beta sheets?

Ans: After computing the propensity of alpha helices and beta sheets, we need to settle for loops.

- Let's see how can we find out the loops using Chou Fasman Algorithm.
- For any jth residue in sequence, we calculate f (Total) = f(j) f(j+1) f(j+2) f(j+3) (tetra peptide).
- If,
- f (Total) > 0.000075.
- The average value for P (turn) > 1.00 in the tetra peptide.

- The averages for the tetra peptide are such P(a-helix) < P(turn) > P(b-sheet),
- It is a Turn!

Q. . How protein structure identified?

Ans: If another protein which has a similar sequence also has its structure known, the
structure of an unknown protein can be predicted based on that similar protein. So, it is then possible to identify unknown protein structures by just examining the homologous protein sequences. Good sequence alignment and identity ensures that homology modelling will give accurate results
Thus, Homology modeling is used to predict structures of proteins having high sequence similarity with other proteins with known structures:

Protein structure prediction:

There are three different strategies for structure prediction
1. Homology Modelling
2. Threading/Fold Recognition
3. Ab Initio Modelling.

Q. What are the properties and role of amino acids?

Ans: Properties:

Amino acids have several properties such as charge
state, polarity and hydrophobicity Amino acids not only have physical and chemical
properties, but also structural properties

Role:

These structural properties are equally important in
giving rise to protein structures Since some amino acids are hydrophobic, they may be employed in forming a stable core in a protein

- Also, chemically inactive amino acids reduce chances
of destabilizing reactions in core

Q. Why folding is important?

Ans: Proteins fold spontaneously. Proteins fold

to achieve thermodynamic stability. Proteins fold to organize themselves for performing functions in cells

Q. write down the terms of PDB file format.

Ans: **HEADER**: contain the brief description of the structure, the date and PDB ID code.

TITLE: title of the structure.

SOURCE: identifies which organism the structure came from.

COMPND: brief deail of the structure.

REVDAT: the date of the last revision.

JRNL: one or more literature reference thet describe the structure.

KEYWDS: list the set of useful words/phrases that describe the structure.

AUTHOR: the scientist depositing the structure.

REMARKS: detail of the experimental method used to determine the structure are contain in this subsection.

Q. define ionization in terms of proteins?

Ans: ionization of protein means to ionize a protein molecule to make it charged by removing an electron from it. Protein ionization is the first step in MS-based proteomics protocols

• Ionization process involves loading of a proton onto a protein or the removal of a proton from the protein

• Ionization results in an increase or decrease of the protein/peptide mass

Two primary techniques used for ionization of protein,

1. Electrospray ionization (ESI)
2. Matrix - assisted laser desorption/ionization (MALDI)

Q. Briefly describe chou fasman algorithm?

Ans: **Chou – Fasman method:**

It is a technique for the prediction of secondary structures in proteins i.e. Alpha Helices, Beta Sheets and Turns is Chou – Fasman technique. The method is based on analyses of the relative frequencies of each amino acid in alpha helices, beta sheets, and turns based on known protein structures solved with X-ray crystallography. From these frequencies a set of probability parameters (in our

handouts, it is propensity table) were derived

- **For the appearance** of each amino acid in each secondary structure type.
- **To predict the probability** that a given sequence of amino acids would form a helix, a beta strand, or a turn in a protein.

Q. How can alpha helix expressed?

Ans: Chou-fasman algorithm (alpha helix):

- For Alpha Helices, 4 contiguous amino acids are required.
- Their Alpha-Helix propensity should be more than 1.0
- Once this propensity falls below 1.0, Alpha-Helix stops.

Q. Describe scores in silico fragmentation?

Ans: Silico fragmentation scoring:

- Count the matches between in silico and in vitro peaks.
- Give an equivalent score to the candidate protein.
- Weigh each of the aforementioned match by the mass error.
- Accumulate the score

Q. what is 3D_1D bowie algorithem?

Ans: Proposed by Bowie et al in 1991

• Converts 3D structure into a 1-D string profile for each structure in the fold library

• Align the target sequence to these profiles 3D-1D methods convert structure and environment information into “profiles”

• Score for each amino acid is computed for each profile

Define Peptide mass finger printing nd shotgun?

Ans: Shotgun proteomics digests the entire protein mix 1st followed by peptide analysis & protein database search

• Peptide mass fingerprinting involves protein separation followed by a single protein’s peptide analysis.

Write steps of protein sequence identification? OR Enlist protein sequence method.

Ans: Mass spectrometers are used to measure the molecular weight of proteins and peptides.

Following steps involve in protein sequence identification,

1. complex protein
2. separation
3. ionization by mass spectrometery
4. fragmentation MS2
5. mass spectra
6. EST Determination
7. filter protein database
8. in silico fragmentation of candidate protein
9.matching of experimental and insilco peak list
10. post translational modification
11. protein score.

Q. write types of RNA ?

ans: **Coding RNAs**

• Coding RNAs as is obvious from their name , code for proetein

• **Non-coding RNAs**

• Non-coding RNAs regulate/assist in the process of translation.Q

Other types of RNA are,

1= m RNA

2=t RNA

3= r RNA

4= miRNA

5= si RNA

Advantage and disadvantage of Ab initio:

Advantages:

• Ab Initio methods can fold any target sequence using only physical atomic properties.

• Predictions are mostly accurate and correctly describe the natural folding process.

Disadvantages:

- Ab initio methods are the very difficult to design (energy function).
- These methods are slow due to the huge possibilities

PAM matrix

- PAM means "Point Accepted Mutations"
- Point accepted mutation is a substitution of one amino acid by another such that the protein functions stays conserved.
- PAM unit is a time in which about 1% of amino acids in a sequence undergo accepted mutations
- PAM matrices are scoring matrices that are useful in computing sequence alignment scores.

STEP TOCOMPUTE PAM MATRICES

1. Align the protein sequence which are 1-PAM Unit diverge.
2. Let Ai,i be the number of times Ai is substituted by Ai.
3. Compute the frequency fi of amino acid Ai.

Then, PAM1=pii= $\frac{Aij}{\sum kAik}$

PAM 'n'= (PAM1)

Steps present in flow chart of homology modeling: - (10 marks)

1. Template recognition and initial alignment.
2. Alignment correction.
3. Backbone generation.
4. Loop modeling.
5. Side-chain modeling.
6. Model optimization.
7. Model validation

How homology modeling is used for knowing the sequence of unknown protein sequence: (10 marks)

If another protein which has a similar sequence also has its structure known, the structure of an unknown protein can be predicted based on that similar protein. So, it is then possible to identify unknown protein structures by just examining

the homologous protein sequences. Good sequence alignment and identity ensures that homology modelling will give accurate results

Thus, Homology modeling is used to predict structures of proteins having high sequence similarity with other proteins with known structures:

How we can get scoring in BUP proteomic

There are two approaches to perform BUP:

- Peptide Mass Fingerprinting.
- Shotgun Proteomics

Score in Protein Fragmentation:-

If we can:

- Measure the mass of fragments using MS.
- Calculate the Protein Fragmentation Techniques theoretical mass of the fragments.

Then, we can award score on the basis of the similarity of experimental and theoretical mass

Protein Fragmentation techniques:

- Electron Capture Dissociation (ECD).
- Electron Transfer Dissociation (ETD).
- Collision Induced Dissociated (CID).

Types of secondary structures of RNA:

1. Single stranded:

- 3'end may fold on to the 5' end.

2. Helices:

- Double stranded RNA helix of stacked base pairs

3. Hairpin loop:

- The loop of the hairpin must at least four bases long to avoid steric hindrance with base-pairing in the stem part of the structure

4. Bulge Loops:

- Bulges, are formed when a double-stranded region cannot form base pairs perfectly

5. Interior loop:

• Interior loops are formed by an asymmetric number of unpaired bases on each side of the loop.

6. Junctions or intersections:

• Junctions include two or more double-stranded regions converging to form a closed structure.

• The unpaired bases appear as a bulge.

Q. How to build scoring matrices?

Ans: To build the Scoring Matrices we analyze the amino acids and nucleotides which are substituted in single gene and protein sequence.

Scoring Matrices have both values +ve and –ve. Positive value for matches and negative value for mismatches. Different type of scoring matrices can be developed based on underlying strategy.

1. Each amino acid have different property.

2. Each amino acid have different frequency.

3. When we compare the sequences they match and mismatch according to their frequency.

Q. Optimal energy function in structural prediction?

Ans: Energy based methods involve evaluating the free energy structures. To compute the RNA sequence for 1' or 2' optimal structure prediction we use Zuker's Algorithm.

RNA Secondary Structure Prediction.

Zuker's Algorithm helps us to compute the stabilizing energies (-ve) and also destabilizing energies (+ve values). And also compute the sum of +ve and –ve energies. All possible 2' structures are generated. The best 2' structure is selected.

Q. Home message of this course?

Ans: 1. Introduce the classical algorithms in bioinformatics

2. Link them to latest developments in the field

3. Evaluate the future applications

Q. Diff. b/w MS and Raw file data?

Ans: **Mass spectrometer** (MS) outputs data with mass/charge ratios & respective ion intensities.

RAW file is a format in which an instrument outputs data in binary form. Several software exist for converting RAW file formats into open software formats. Each open format has its own unique advantages. mzXML and MGF formats are most frequently used

Q. Why we need make protein structure when we have X-Ray and NMR?

Ans: There are thousand of sequences and lesser structures are available. Reason is that some proteins cannot be crystalized or it may not be soluble so we have need to make protein structure althou we have x- ray crystallography.

Q. Amino acid with 3 codon?

Ans: lysine has 3 codon. (Lie = AUU, AUG, AUA)

Q. How + and – energies released in molecule?

Ans: The stabilizing energy associated with stacking base pairs in a double-stranded region is (-ve) energy.

• The destabilizing influence of unpaired regions is (+ve) energy.

Q. How back bone of protein form?

Ans: C termini in a protein backbone C-Alphas can be used to construct the backbone of a protein towards its visualization. Proteins have Carbon and Nitrogen in their backbone

Q. which factor consider during N-J trace back?

Ans: 'Trace back' strategy is used to recover the optimal structure, there can be multiple trace backs. Each trace back can be used to construct an RNA secondary structure. NJ Select the trace back with the highest number of coupled nucleotides.

Q. Top down proteomics?

Ans: for a direct measurement of protein mass

• Solution: TDP

1. Samples containing the protein mixture from cells or tissue are obtained
2. The entire protein mix is analyzed for protein masses (MS1)
3. The mass list thus obtained has masses of all intact proteins

4. Note that proteins typically carry post-translational modifications and TDP caters for their mass measurement as well
5. After MS1, one protein is selected at a time and fragmented to obtain its peptides (MS2)
6. Several protein fragmentation techniques exist for fragmenting proteins
7. Fragmentation products i.e. peptides are measured for their mass (MS2)
8. The process is repeated (MS'n')

Conclusions:

• TDP measures the mass of intact proteins
• Mass of post translational modifications can be easily measured.

Q. Formula of scoring matrices?

Ans: pii= $\frac{Aij}{\sum kAik}$

Q. What are forces used in amino acids to fold a protein?

Ans: Electrostatic interactions
• van der Waals interactions
• Hydrogen bonds
• Hydrophobic interactions

Q. How OH of RNA make it less stable than DNA?

Ans: The two hydroxyl groups make RNA less stable than DNA because it is more prone to hydrolysis Hence, RNA has a short life span and is more prone to degradation.

REMEMBER ME IN YOUR PRAYERS

*BEST OF	LUCK*