Secondary structure prediction

1. Secondary structures are stable local conformations of a polypeptide chain
2. They are important in maintaining 3D structure of the protein
3. 50% of residues in a protein fold into either
    a. Alpha helix
        i. Spiral structure
        ii. 3.6 amino acid per turn
        iii. Stabilized using Hydrogen bonds between residues I and i+4
        iv. Prolines occur only at the end position of alpha helices
    b. Beta sheet
        i. Consists of two or more Beta strands having extended zig-zag conformation
        ii. Stabilized by hydrogen bonds between residues of adjacent strands
        iii. May have long range interactions on primary structure level
        iv. Surface Beta strands show alternating hydrophobic and hydrophilic residues
        v. Buried Beta strands contain mainly hydrophobic residues
4. Secondary structure prediction refers to prediction of the conformational state of each amino acid residue of a protein sequence based on their three possible states
    a. Helix [H]
    b. Strand [E]
    c. Coil [C]
5. Prediction is based on fact that secondary structures have regular arrangement of amino acids
6. This structural regularity serves as foundation of prediction algorithms
7. Uses
    a. Classification of proteins for separation of domains and motifs
    b. Secondary structures are much more conserved than sequences during evolution

Ab initio-based methods:

1. Predicts secondary structure based on single query sequence
2. Each amino acid is given a propensity score
3. Propensity scores are derived from known crystal structures

Homology based methods:

1. Third generation was developed in late 1990s
2. Combines ab initio secondary structure prediction of individual sequences with alignment information from multiple homologous sequences

Chou & Fasman Secondary structure prediction server (first generation)

1. Online secondary structure prediction server
2. Predicts regions of secondary structure from the amino acids such as
    a. Alpha helix
    b. Beta sheet
    c. Turns
3. Output is displayed in linear sequential graphical view based on probability of occurrence of
    a. Alpha helix
    b. Beta sheet
    c. Turns
4. Method implemented in Chou-Fasman is based on analysis of each amino acid in alpha helixes, beta sheets and turns based on known protein structures
5. CFSSP is freely accessible via ExPASy or BioGem tools

GOR IV (second generation)

1. Based on information theory
2. Developed by Garnier, Osguthorpe and Robson (hence GOR)
3. Used all possible pair frequencies withing 17 amino acid residues
4. Cross validates on a database of 267 proteins
5. Has accuracy of 64.4% for Q3 prediction. (Q3 stands for average of each Qi where i stands for alpha helix, beta sheet, loop)
6. Program gives 2 outputs
    a. Sequence and predicted secondary structure represented in rows showing
        i. Helix
        ii. Extended
        iii. Coil
    b. Second give probability values for each secondary structure at each amino acid position

PSIPRED (third generation)

1. Used two feed-forward neural networks
2. Performs analysis on output obtained from PSI-BLAST
3. It uses a stringent cross validation procedure
4. Has Q3 accuracy of 76.5%
5. Has a Java front-end application called PSIPREDView which interprest results from PSIPRED

Introduction to protein Classification

1. Classification of protein into
    a. sequence
    b. structure groups
2. Structure comparison allows identification of relationships among structures
3. To establish hierarchical relationships among protein structures
4. Provide comprehensive and evolutionary view of known structures
5. Once classification system is established new protein can find a place in a category and its functions can be better understood
6. First step is to remove redundancy from databases
    a. Majority structures are solved at different resolutions or associate with different ligands or with single residue mutations
    b. This makes most structures repetitive and redundant
7. Redundancy can be removed by
    a. Selecting representatives via a sequence-alignment based approach
    b. Proteins are composed of multiple domains and need to be separated into individual domains
        i. These domains need to be subdivided before a sensible structural comparison can be carried out
        ii. Domain identification can done
            1. Manually
            2. Special algorithms for domain recognition
8. Once separation is done, structure comparison is done at domain level
9. Last step can be done manually or via automated means or a combination of both

CATH

1. Developed in 1997

2. Provides up to date and systematic structural classification of protein 3D structures
3. Core Data Resources within ELIXIR
4. Semi-automated process to split 3D structures into individual domains
5. Clusters domains into superfamilies where there is sufficient evidence of evolutionary ancestry
6. Assigns domains for unknown protein sequences
7. Both CATH and Gene3D provide detailed structural domain assignment and annotation
8. To predict domain
    a. Use a set of representative domains to seed a set of protein alignments
    b. Alignments are converted to Hidden Markov Models
    c. HMMs are used to identify closely related domains within the sequence from UniProt and ENSEMBL
9. Provides comprehensive structure-based domain superfamily assignment for over 82 million protein sequences
10. Domains classified into hierarchical levels
    a. Class C
    b. Architecture A
    c. Topology T
    d. Homologous superfamilies H
11. CATH further subclassifies homologous superfamilies into Functional Families (FunFams)
    a. Made on the basis of specificity-determining positions
    b. Calculates a functional coherence index to determine functionally coherent alignments
12. FunFams are more functionally coherent than other domain-based approaches
13. For each FunFams sequence alignment, high quality GO annotations are provided

SCOPe

1. Structural Classification of proteins – extended (SCOPe)
2. Provides accurate, detailed and comprehensive description of structural and evolutionary relationships amongst majority of the known protein structures

3. Structures are divided into domains using manual curation and highly precise automated methods
4. Heirarchy of SCOPe
    a. Species
        i. Distinct protein and its naturally occurring or artificially created variants
    b. Protein
        i. Group of similar sequences that perform same function but originate from different biological species
    c. Family
        i. Similar sequences but different functions
    d. Superfamily
        i. Bridges protein familyies with common functions and structural features thought to be from common ancestor
    e. Folds
        i. Structurally similar superfamilies
    f. Classes
        i. Mainly based on secondary structure content and organization
5. Members of superfamily are considered to have common ancestral origin although family relationships are considered distant
6. Classes is the highest level of hierarchy since it distinguishes secondary structures such as all alpha, all beta, alpha and beta, etc.

Tertiary structure prediction

1. Aims to predict three-dimensional shape of protein molecules by describing spatial disposition of each atom
2. There are methods to resolve molecular structure with high precision but they are time and resource consuming
3. Computation based software techniques can predict tertiary structure of protein with acceptable precision with high efficiency
4. Currently solving a protein takes 1 to 3 years.
5. Certain proteins are extremely difficult to resolve by Xray and NMR techniques
6. Sequence data for many important proteins is available but the 3D structure is unknown
7. Hence full understanding of their biological roles cannot be studied

8. It is necessary to obtain approximate protein structure through computer modelling
9. Uses
    a. Rational design of biochemical experiments
        i. Site directed mutagenesis
        ii. Protein stability
        iii. Function analysis
    b. Help explain experimental results
    c. In short helps advance understanding of protein function

Method 3D structure prediction

Homology modelling

1. Predicts structure based on sequence homology with known structures
2. Known as comparative modelling
3. Based on principle that if two proteins share high enough sequence similarity, they are likely to have similar 3D structure
4. If one structure is known it can be copied to unknown sequence
5. Produces all atom model
6. Six steps
    a. Template selection
        i. Identification of homologous sequences to be used as templates for modelling
    b. Alignment of target and template
    c. Framework building
        i. For target protein consisting of main chain atoms
    d. Model building
        i. Addition of side chain atoms and loops
    e. Refine and optimize
        i. Done according to energy criteria
    f. Evaluating overall quality
    g. Alignment and model building is repeated until satisfactory result is obtained

Modeller

1. Computer program for comparative protein structure modelling
2. Input is alignment sequences to be modelled
3. Automatically calculates model containing all non hydrogen atoms

4. Apart from model building it can perform auxiliary tasks
    a. Fold assignment
    b. Alignment of two protein sequences
    c. Multiple alignment
    d. Calculation of phylogenetic trees
    e. De novo modelling

I-TASSER (Threading and folding recognition)

1. Online platform
2. Implements I-Tasser based algorithms for protein struc and func predictions
3. Automatically generate high quality model predictions
4. Threading and fold recognition predicts the structural fold of unknown protein by fitting sequence in a database and selecting best fitting fold
5. This approach can identify structurally similar protein even without detectable sequence similarity
6. C-Score: Confidence score for estimating quality of predicted models
7. TM-Score: Scale for measuring structural similarity between two structures

Robetta (Ab initio protein structural prediction)

1. Provides automated tools for protein structure prediction and analysis
2. Sequences are parsed into domains and models are generated using comparative or de novo prediction methods
3. If match is found comparative modelling is performed if match is not found de novo modelling is performed
4. Ab initio method attempts to produce all-atom protein models based on sequence information along without aid of known structures

Saves server

Errat:

1. Method for differentiating between correctly and incorrectly determined regions of protein structures
2. Errat is a program for verifying protein structures determined by crystallography
3. Confidence level of 95% is yellow
4. Confidence level of 99% is red

Verify3D:

1. Used precomputed database of 18 environmental profiles, compiled in high resolution to ass the quality of the protein model
2. Result is 2d graph showing folding quality of each residue of the protein structure
3. Residues with score below 0 are considered unfavourable

Proves:

1. Calculates volumes of atoms in the macromolecule
2. Calculates a statistical Zscore deviation from highly refined and resolved structures

Whatcheck:

1. Does checking of many stereochemical parameters
2. Has many functions
    a. Checking planarity
    b. Collions
    c. Symmetry
    d. Proline puckering
    e. Anomalous bond angles
    f. Bond lengths

Procheck:

1. Check general physicochemical parameters
2. Chirality
3. Bond lengths
4. Bond angles

Cryst:

1. Searches PDB entries

Multiple verification methods are used because no method is superior than the other

RASMOL

1. Computer program for visualization of biological macromolecule structures
2. Developed by Roger Sayle in 90s

3. Important tool since it can run on modest hardware efficiently
4. Ground breaking educational tool
5. Can represent
    a. Wireframe
    b. Culinder
    c. Stick bonds
    d. Alpha carbon trace
    e. Space filling spheres
    f. Ribbons
6. Imp commands
    a. Select <type of bond>
    b. Select <element>
    c. Label <what to label the selection>
    d. Label false (to label the previous selection)

PyMOL

1. Opensource molecular visualization system
2. Cross platform
3. Widely used
4. Can visualize
    a. Proteins
    b. Nucleic acids
    c. Small molecules
    d. Electron densities
    e. Surfaces
    f. Trajectories
5. Can represent
    a. Ribbons
    b. Cartoons
    c. Dots
    d. Surfaces
    e. Spheres
    f. Sticks
    g. Lines
6. Many plugins are available

castp

1. Geometric and topological properties of protein structures, including surface pockets, interior cavities and cross channels, are of fundamental importance for proteins to carry out their functions.
2. Computed Atlas of Surface Topography of proteins (CASTp) is a web server that provides online services for locating, delineating and measuring these geometric and topological properties of protein structures.
3. **imprints** of the negative **volumes** of **pockets**, **cavities** and channels,
4. **topographic features** of biological assemblies in the Protein Data Bank,
5. **improved visualization** of protein structures and pockets,
6. **more intuitive structural** and **annotated information**, including information of secondary structure, functional sites, variant sites and other annotations of protein residues

netncglyc

(To predict binding pocket for Glycosylation sites)

1. NetNGlyc 4.0 Server can be used to predict the glycosylation sites of proteins that enables proteome-wide discovery of O-glycan sites using ‘bottom-up’ ETD-based mass spectrometric analysis

2. This information can be used by researchers for a wide range of studies, including

3. investigations of signaling receptors,

4. discoveries of cancer therapeutics,

5. understanding of mechanism of drug actions,

6. studies of immune disorder diseases,

7. analysis of protein–nanoparticle interactions,

8. inference of protein functions

netphos

1. NetPhos - 3.1 server can be used to predict the phosphorylation sites of

proteins.

2. Protein phosphorylation at serine, threonine or tyrosine residues affects a multitude of cellular signaling processes.
3. Phosphorylation sites predicted by neural networks. NetPhos - 3.1 server (Generic phosphorylation sites in eukaryotic proteins) is used.
4. server predicts serine, threonine or tyrosine phosphorylation sites in eukaryotic proteins using ensembles of neural networks
5. This information can be used by researchers for a wide range of studies, including
    - investigations of signaling receptors,
    - discoveries of cancer therapeutics,
    - understanding of mechanism of drug actions,
    - studies of immune disorder diseases,
    - analysis of protein–nanoparticle interactions,
    - inference of protein functions

Vast +

1. detection of similarities between protein 3D structures for detection of homologous relationships, the classification of protein families and functional inference.
2. VAST+, an extension to the existing VAST service, which summarizes and presents structural similarity on the level of biological assemblies or macromolecular complexes.
3. VAST+ simplifies structure neighboring results and shows, for macromolecular complexes tracked in MMDB, lists of similar complexes ranked by the extent of similarity.
4. VAST+ replaces the previous VAST service as the default presentation of structure neighboring data in NCBI's **Entrez query and retrieval system.**

Dali

The Dali server is a network service for comparing protein structures in 3D. You submit the query protein structure and Dali compares them against those in the Protein Data Bank (PDB).

comparing 3D structures may reveal biologically interesting similarities that are not detectable by comparing sequences.

User can perform three types of database searches:

1. **Heuristic PDB search** - compares one query structure against those in the Protein Data Bank
2. **Exhaustive PDB25 search** - compares one query structure against a representative subset of the Protein Data Bank
3. **Hierarchical AF-DB search** - compares one query structure against a species subset of the AlphaFold Database

Tssw-

1. distinguishes promoter sequences from non-promoter sequences
2. Algorithm predicts potential transcription start positions by linear discriminant function combining characteristics describing functional motifs and oligonucleotide composition of these sites

Bprom

1. bacterial promoter
2. It uses a linear discriminant function
3. combined with signal and content information such as consensus promoter sequence and oligonucleotide composition of the promoter sites
4. This program first predicts a given sequence for bacterial operon structures by using an intergenic distance of 100 bp as basis for

distinguishing genes to be in an operon.

5. the program is most effectively used when about 200 bp of upstream sequence of the first gene of an operon is supplied as input to increase specificity.

FGENESB

1. To predict bacterial operon (a cluster of genes that are transcribed together to give a single messenger RNA (mRNA) molecule, which therefore encodes multiple proteins)
2. The program is specifically trained for bacterial sequences.
3. It uses the Vertibi algorithm to find an optimal match for the query sequence with the intrinsic model.
4. A **linear discriminant analysis (LDA**) is used to further distinguish coding signals from noncoding signals.

FGENES

1. is a web-based program that uses LDA to determine whether a signal is an exon.
2. In addition to FGENES, there are many variants of the program. eg FGENESH making use of HMMs
3. Some programs, such as FGENESH+, combine both ab initio and similarity-based approaches.

ORF

1. ORF finder can used to predict open reading frames in the genome.
2. The program returns the range of each ORF, along with its protein translation.
3. Use ORF finder to search newly sequenced DNA for potential protein encoding segments, verify predicted protein using BLAST
4. This information of long ORFs are often used, along with other evidence, to initially identify candidate protein-coding regions or functional RNA_coding regions in a DNA sequence.

5. Small Open Reading Frames (small ORFs/sORFs/smORFs) are important sources of putative peptides previously dismissed as being non-functional or junk DNA.

whole-genome sequencing

1. Easy to generate de novo draft genome sequences for any organism of choice
2. a genome project requires careful planning with respect to the organism involved and the intended quality of the genome draft.
3. Genome projects employ DNA sequencing, mapping, and computational technologies to understand molecular/cellular mechanisms, gene repertoires, genome architecture, and evolution
4. Such technical advantages and established recommendations and strategies have been widely applied in humans, terrestrial animals, and plants and crops we could do aquatic species too but aquaculture applications are slower as compared to the rest.
5. DNA sequencing is now routinely carried out using the Sanger method. This involves the use of DNA polymerases to synthesize DNA chains of varying lengths.
6. The DNA synthesis is stopped by adding dideoxynucleotides. The dideoxynucleotides are labeled with fluorescent dyes, which terminate the DNA synthesis at positions containing all four bases, resulting in nested fragments that vary in length by a singlebase.
7. When the labeled DNA is subjected to electrophoresis, the banding patterns in the gel reveal the DNA sequence.
8. DNA sequences are read by a computer program that assigns bases for each peak in a chromatogram. This process is called base calling. automated base calling has errors so its often required to correct the sequence calls.
9. two major strategies for whole genome sequencing: the **shotgun approach** and the **hierarchical approach**

10. **SGA: sequencing many overlapping DNA fragments in parallel and then using a computer to assemble the small fragments into larger contigs and, eventually, chromosomes** This is designed to minimize sequencing errors and ensure correct assembly of a contiguous(adjoining) sequence
11. **HA:s** similar to the shotgun approach, but on a smaller scale **begins by first generating a physical map. The overlapping clones that define the map are then shotgun cloned and sequenced**
12. Although the approach has been successfully employed in sequencing small microbial genomes, for a complex eukaryotic genome that contains high levels of repetitive sequences, such as the human genome, the full shotgun approach becomes less accurate and tends to leave more "holes" in the final assembled sequence than the hierarchical approach.
13. Current genome sequencing of large organisms often uses a combination of both approaches.

Genome seq assembly:

1. The first step toward genome assembly is to derive base calls and assign associated quality scores.
2. The next step is to assemble the sequence reads into contiguous sequences. This step includes identifying overlaps between sequence fragments, assigning the order of the fragments and deriving a consensus of an overall sequence.
3. To assemble a whole genome sequence, these short fragments are joined to form larger fragments after removing overlaps. These longer, merged sequences are termed contigs, which are usually 5,000 to 10,000 bases long.
4. Assembling all shotgun fragments into a full genome is a computationally very challenging step.
5. There are a variety of programs available for processing the raw sequence data

GENOME ANNOTATION:

1. DNA annotation or genome annotation is the process of identifying the locations of genes and all of the coding regions in a genome and determining what those genes do. An annotation is a note added by way of explanation or commentary. Once a genome is sequenced, it needs to be annotated to make sense of it.

GENE ONTOLOGY

1. The Gene Ontology (GO) project is a major bioinformatics initiative to develop a computational representation of our evolving knowledge of how genes encode
2. There are three different ontologies: cellular components, biological processes and molecular functions
3. **Publishing the genome:** ensembl, ncbi provides opportunity to upload gene drafts

UCSC

1. UCSC Genome Bioinformatics Group and external collaboratord display
2. gene predictions,
3. mRNA and expressed sequence tag alignments,
4. simple nucleotide polymorphisms,
5. expression and regulatory data,
6. phenotype and variation data, and
7. pairwise and multiple_species comparative genomics data.
8. All information relevant to a region is presented in one window,
9. facilitating biological analysis and interpretation.

Ensembl

1. is a genome browser for vertebrate genomes that supports research in
2. comparative genomics,
3. evolution, sequence variation and transcriptional regulation.
4. Ensembl annotate genes,
5. computes multiple alignments,
6. predicts regulatory function and collects disease data.
7. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.