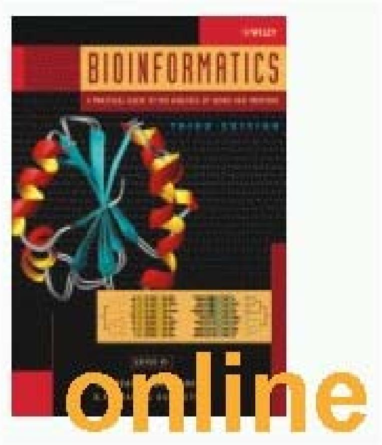
4. Predictive Methods Using Protein Sequences (1)

薛 佑 玲 Yow-Ling Shiue
國立中山大學生物醫學研究所
✉ ylshiue@mail.nsysu.edu.tw



Select a Chapter: Chapter 8

Chapter 8: Predictive Methods Using Protein Sequences

- Sample Data for Problem Sets
- Internet Resources

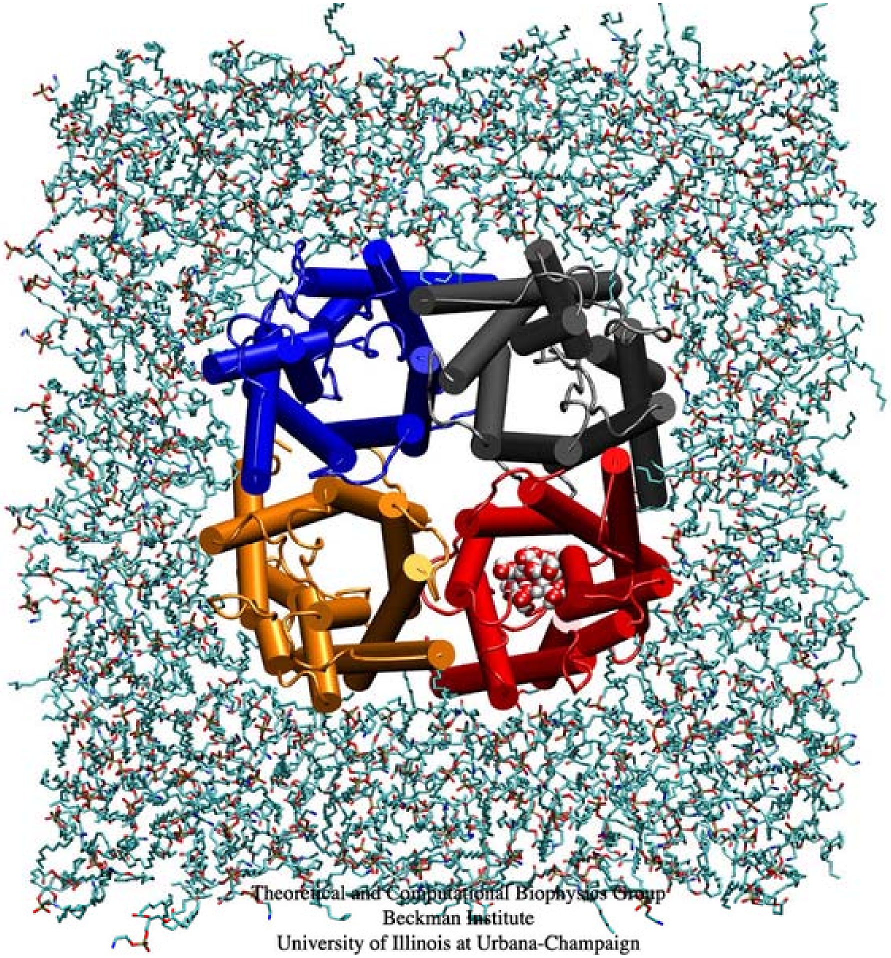
Introduction (1)

- **The central dogma**
  - Linear flow of information
    - **DNA ⇨ RNA ⇨ protein**
    - But not conversely
- **Proteins**
  - The function (or in a case of **disease**, the **malfunction**) of each protein is encrypted in its amino acid sequence
    - **Easy to do**
      - DNA & protein sequencing
    - **Difficult to do**
      - Protein **structure** & **function** experiments

Introduction (2)

- Gold Genomes Database
  - http://www.genomesonline.org/
    - Sequenced and sequencing genome projects

- Most protein sequences
  - No experimental annotations
  - To develop **computational tools** to decipher the information encoded in **protein sequences** ⇨ to predict their **structure** & **function**

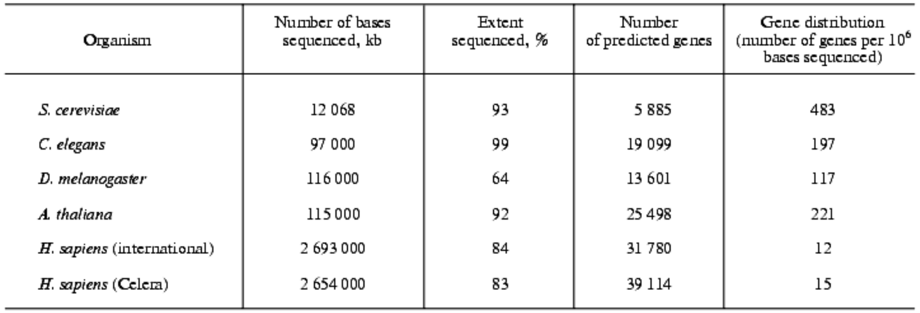


Theoretical and Computational Biophysics Group
Beckman Institute
University of Illinois at Urbana-Champaign

Introduction (3)

- DNA sequences include **additional information** that could not be extracted from the corresponding protein sequence
  - **Instructions** for the **control** & **regulation** of protein expression

- Prediction of **gene number** in completed sequencing genomes were not even close
  - E.g., Drosophila: 13,000 (2000) ⇨ 19,000 (now)



Organism	Number of bases sequenced, kb	Extent sequenced, %	Number of predicted genes	Gene distribution (number of genes per 10^6 bases sequenced)
S. cerevisiae	12 068	93	5 885	483
C. elegans	97 000	99	19 099	197
D. melanogaster	116 000	64	13 601	117
A. thaliana	115 000	92	25 498	221
H. sapiens (international)	2 693 000	84	31 780	12
H. sapiens (Celera)	2 654 000	83	39 114	15

- Study on **protein sequences** can by pass the problems inherent in analyzing only nucleotide sequences

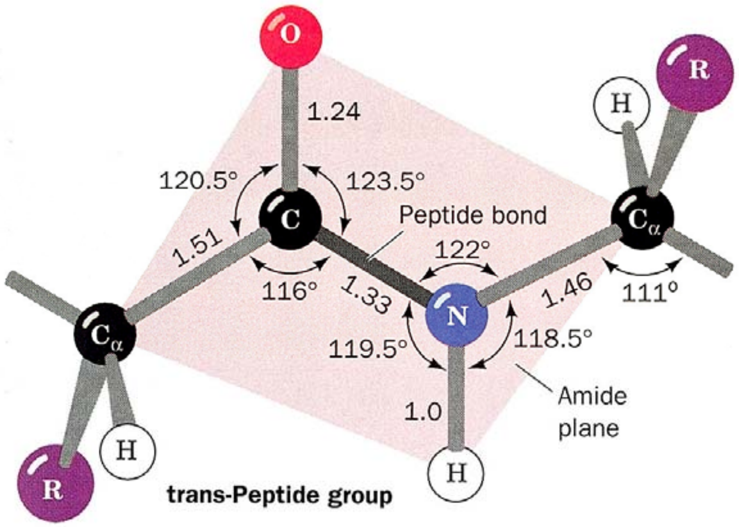
Introduction (4)

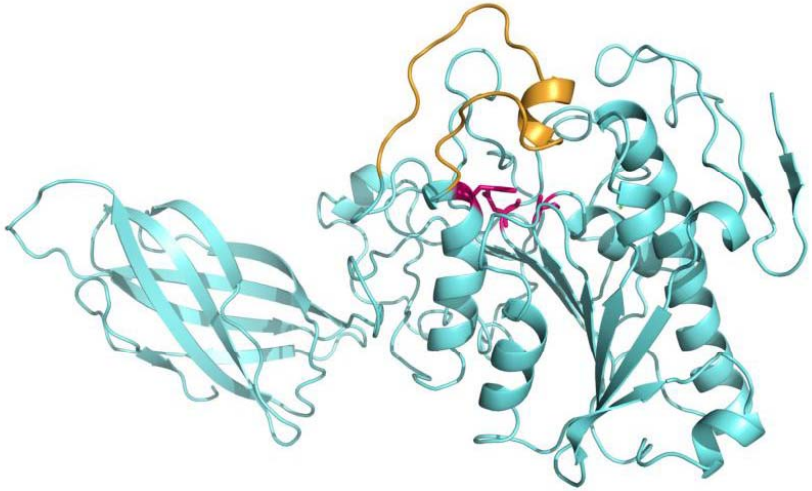
- **Known protein sequences**
  - To analyze the information **relating directly** to a protein's structure & function

- **>300 different methods** to predict 2nd structure of proteins
  - **Only a very few** are geared toward predicting **functional classes**

Predicting Features of Individual Residues (1)

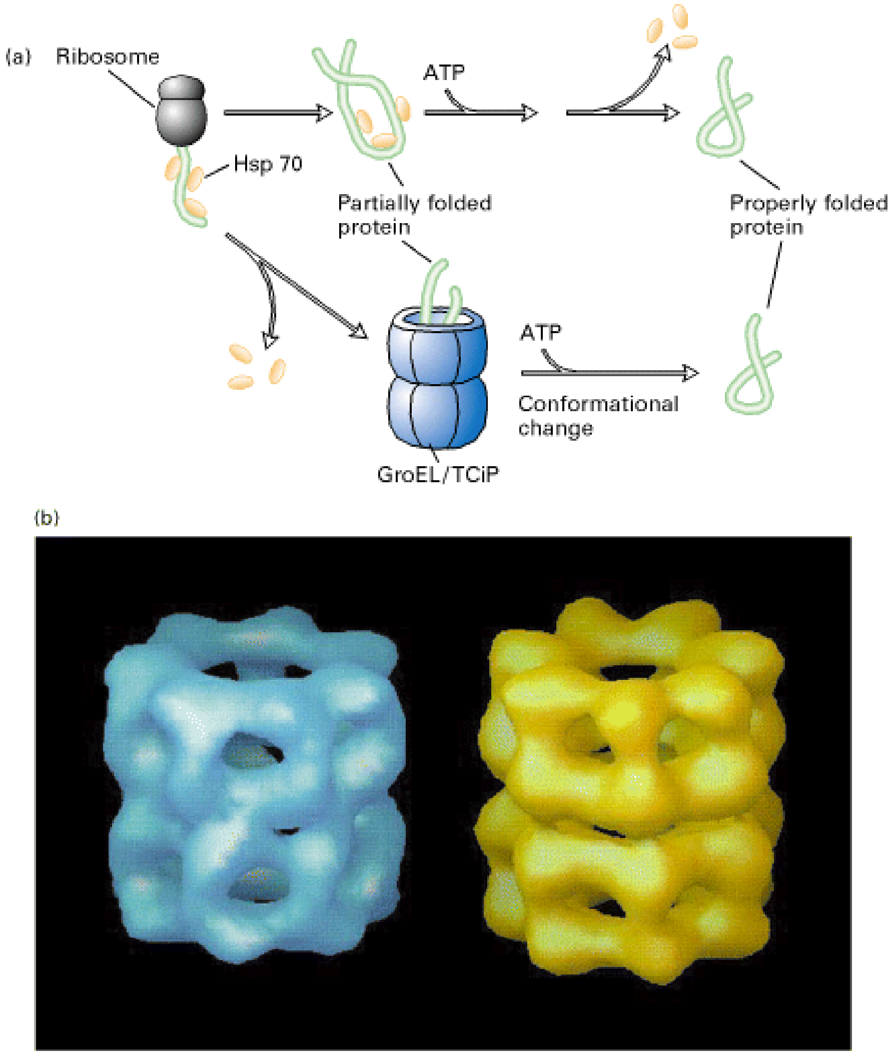
- Residues (shapeless) + **peptide bonds** = chain → → → **3D structure** (in a process of not fully understood)
  - But, structure is determined for the most part, **by sequence alone**





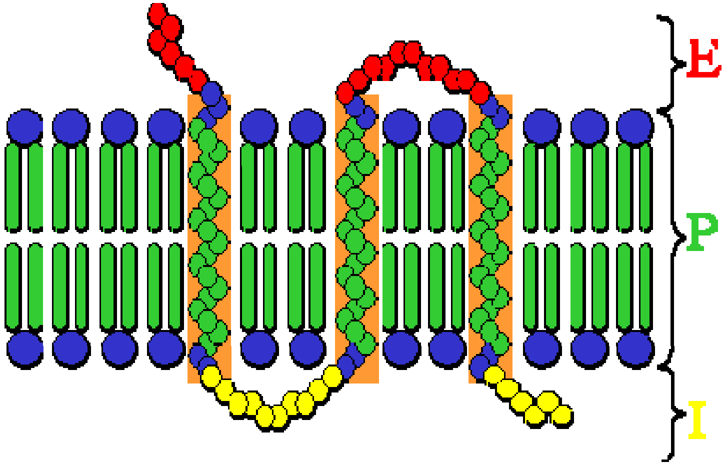
Predicting Features of Individual Residues (2)

- Many proteins do not fold properly **in the test tubes**, they require
  - **Additional information** required that is **not available** in the **sequence** or
  - Required assistance from **other proteins** to adopt their native structures
    - **Chaperonins**
- **Assumptions**
  - Native 3D structure: **energetic minimum**



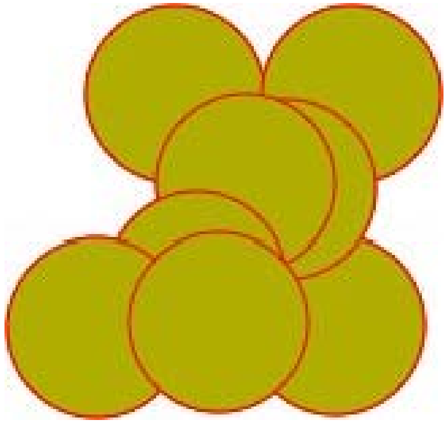
Predicting Features of Individual Residues (3)

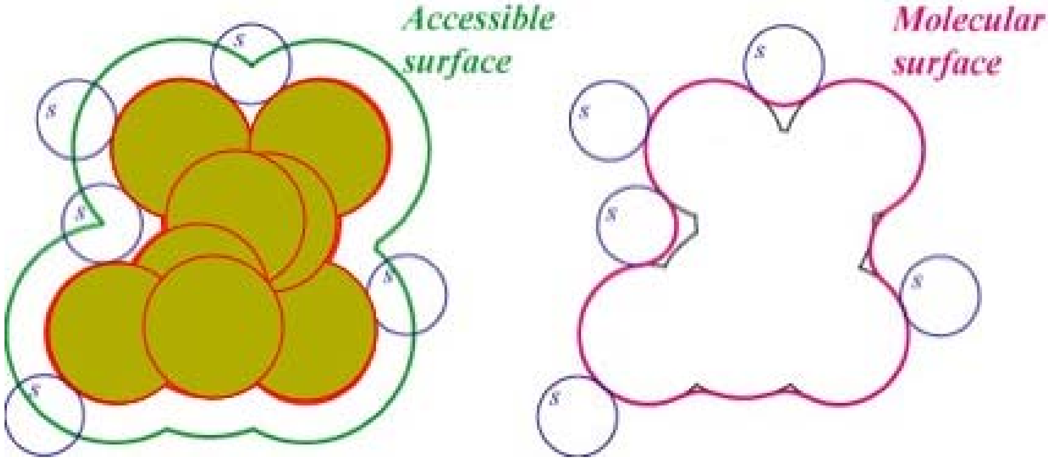
- Early pioneering **structure prediction methods**, the **structural characteristics** of individual residues
  1) Simplified
     - It breaks the problem of structure prediction down into smaller elements
     - Easy to handle
  2) To identify residues with certain **features, e.g., membrane- or solvent-accessible residues**



Solvent Accessibility

Each atom can be modeled as a Van der Waals sphere in three dimensions. The union of the spheres give the molecular surface





Not all molecular surface is accessible to solvent due to the existence of small cavities. **Rolling a solvent ball over the Van der Waals spheres traces out the surface area experienced by the solvent.** **Solvent accessible surface** area (SASA) is a very important measure for quantitatively determining the behavior and interaction tendencies of a protein

Secondary Structure Prediction (1)

- Protein **sequence** = **primary structure**
- Short stretches of residues tend to form **local structures** = **secondary structure**
- Overall three dimension structure of a protein chain = **tertiary structure**
  - A **topological** organization & spatial packing of **smaller local structures**

- The **driving forces** behind the formation of **secondary structures**
  - A complex combination of **local** & **global** forces

Local Forces

- Acting **between residues**, or **between** the **residue** & the **backbone** of the protein
  - Repulsion **between hydrophobic** side chains of some amino acids & the **hydrophilic** backbone of the protein chain
  - The **interaction** between **side chains** & the surrounding solvent (Pauling et al. 1951)
  - The subcellular environment of the **protein** surrounding can also affect its secondary structure

Global Forces

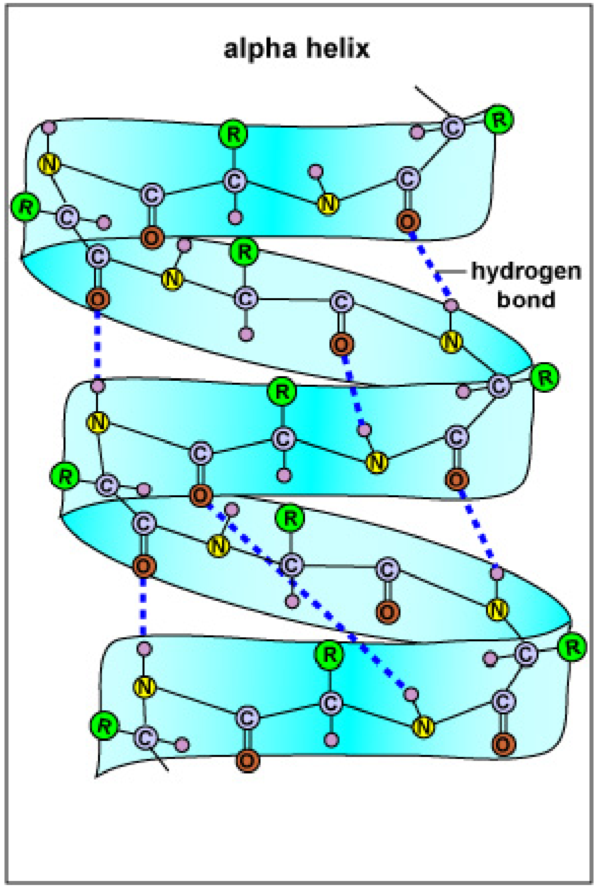
- **Global forces**
  - Forces exerted by **other**, **more distant parts** of the structure
    - Repulsing or attracting each other in **complex manners**
- To **predict** secondary structure
  - **Challenging**
- **Three types**
  - **Helices**, **strands**, **others** (non-regular or **loop regions**)

Helix

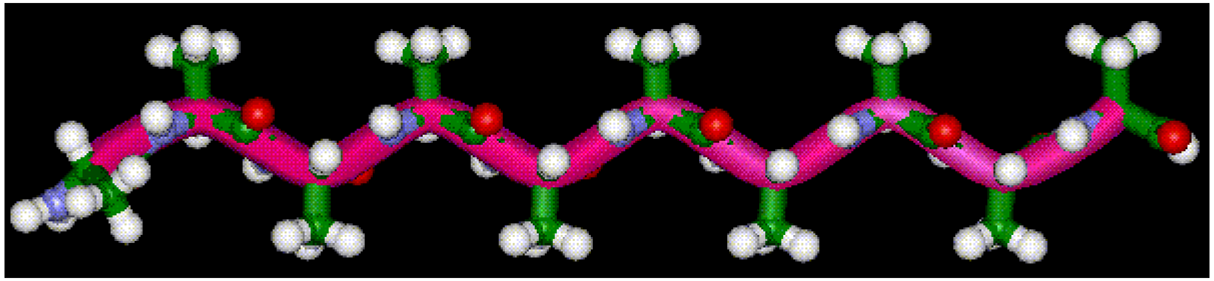
- A corkscrew-like spiral of **the backbone** with the side chains projecting outward in different angles
  - Several subtypes

- The most common type
  - ***Alpha-helix***
    - 3.6 residues in each turn of the spiral

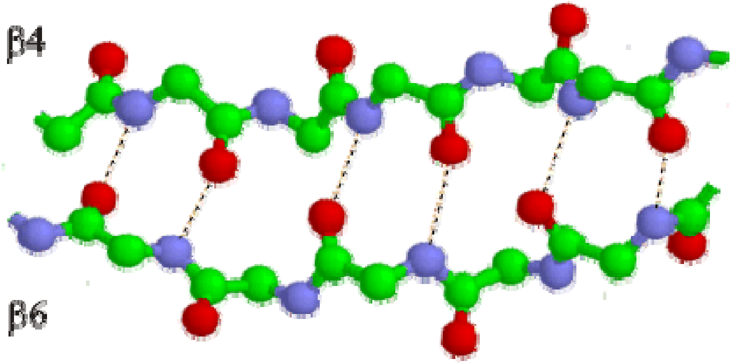
- **Hydrogen bonds** between **the carbonyl (CO) group** of one amino acid & **the amino (NH) group** of the C-terminal, 4th amino acid



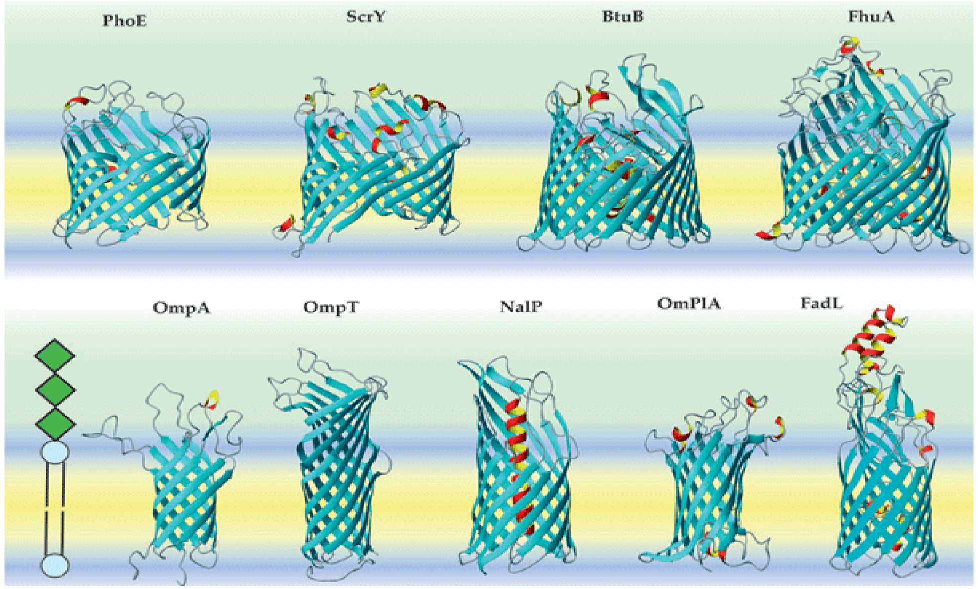
Strand



- Most common: **beta-strand**
- **Two or more** stretches of beta-strands often **interact** with each other, **through hydrogen bonds**
  - >2 different types of strands



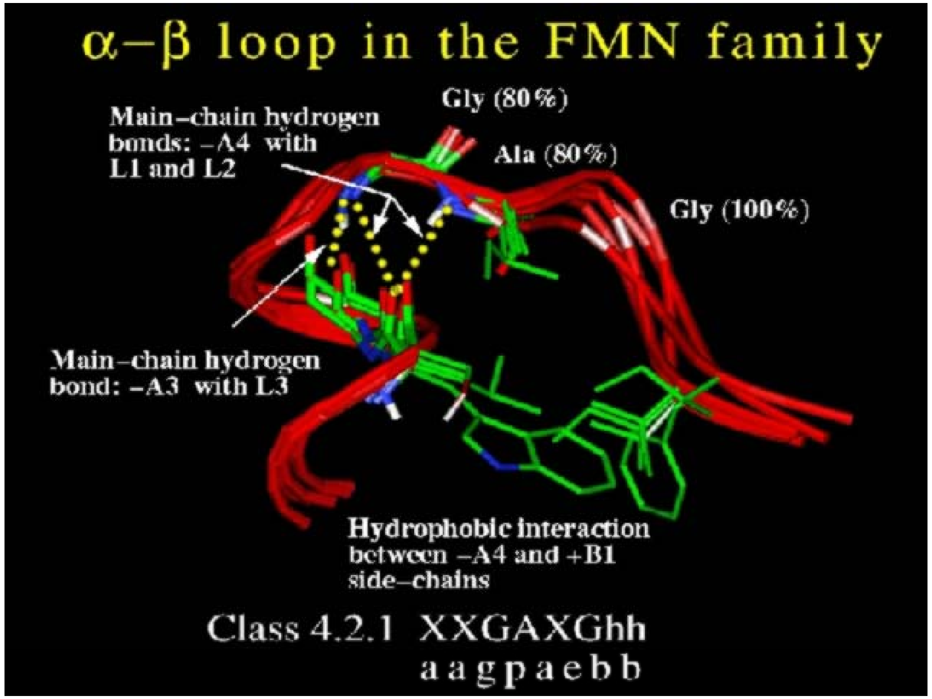
RNase A: single chain, **anti-parallel strands**



Some known structures of **β-barrel** membrane proteins

Loop

- The local structures, **not** helix & not strand
  - A few types of loops


α–β loop in the FMN family
Gly (80%)
Main–chain hydrogen bonds: –A4 with L1 and L2
Ala (80%)
Gly (100%)
Main–chain hydrogen bond: –A3 with L3
Hydrophobic interaction between –A4 and +B1 side–chains
Class 4.2.1 XXGAXGhh
aagpaebb

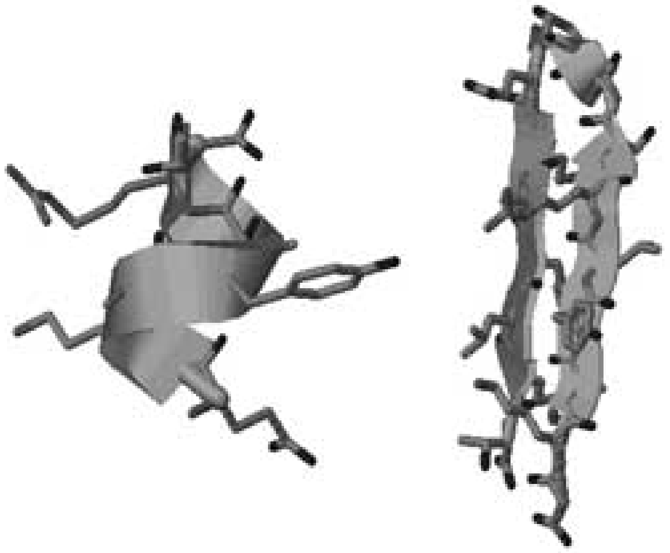
Analysis of Secondary Structure (1)

- Different amino acids tend to **“prefer”** being contained in **different secondary structure**
  - Relating secondary structure to amino acid propensities reemphasizes the importance of **local sequence environment** in the formation of secondary structure

- Most **prediction methods**
  - Predominantly rely on **local sequence information** in making their **determinations**
  - But **non-local forces** can be crucial

Analysis of Secondary Structure (2)

- **Extreme example**
  - The **Chameleon protein** (Minor & Kim 1996)
    - A stretch of **11 consecutive amino acid residues** adopts a **helical** structure in one region & a **strand** at another
      - **Local information is not enough**



Prediction Methods (1)

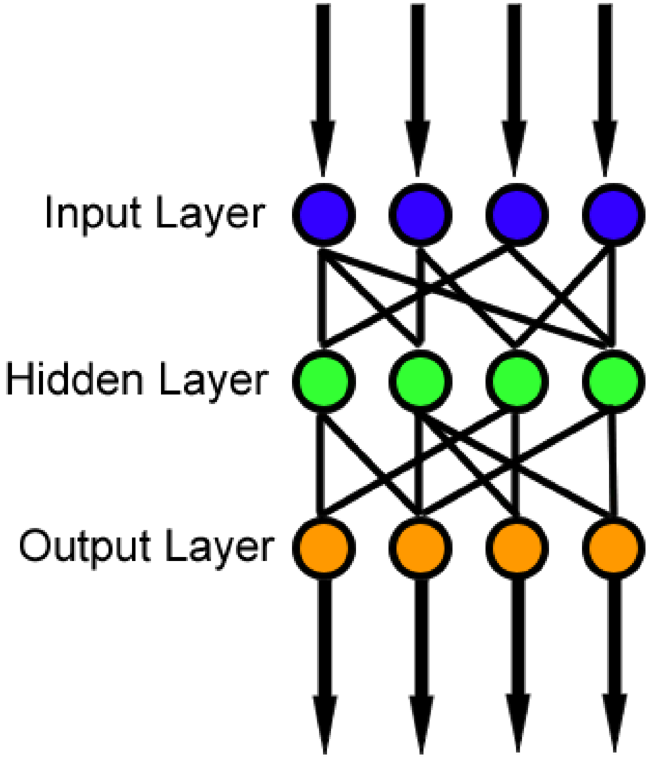
- **PHDsec** (Rose et al. 1994; 1996) & **PROFsec** (Rost et al. 2003)
  - Part of the PredictProtein (Rost et al. 2003) service for sequence analysis & structure prediction
  - Both shares the same **basic approach**
  - **PROFsec** = an improved & overly complicated version of PHDsec
- *Algorithms*
  - **Machine learning**: learn from **a large set of examples**
    - **The training set**
      - The **implicit rules & principles underlying** a certain phenomenon

Machine-learning Algorithms (1)

- The success of machine-learning algorithms dependent on
  - The careful choice of **biologically based features used for training**
    - E.g., the **residues surrounding** a particular residue i (i.e., the residue of i-n, … i+n) have a tremendous effect on the secondary structure of residue i
  - Input = **surrounding residues** + target residues

Machine-learning Algorithms (2)

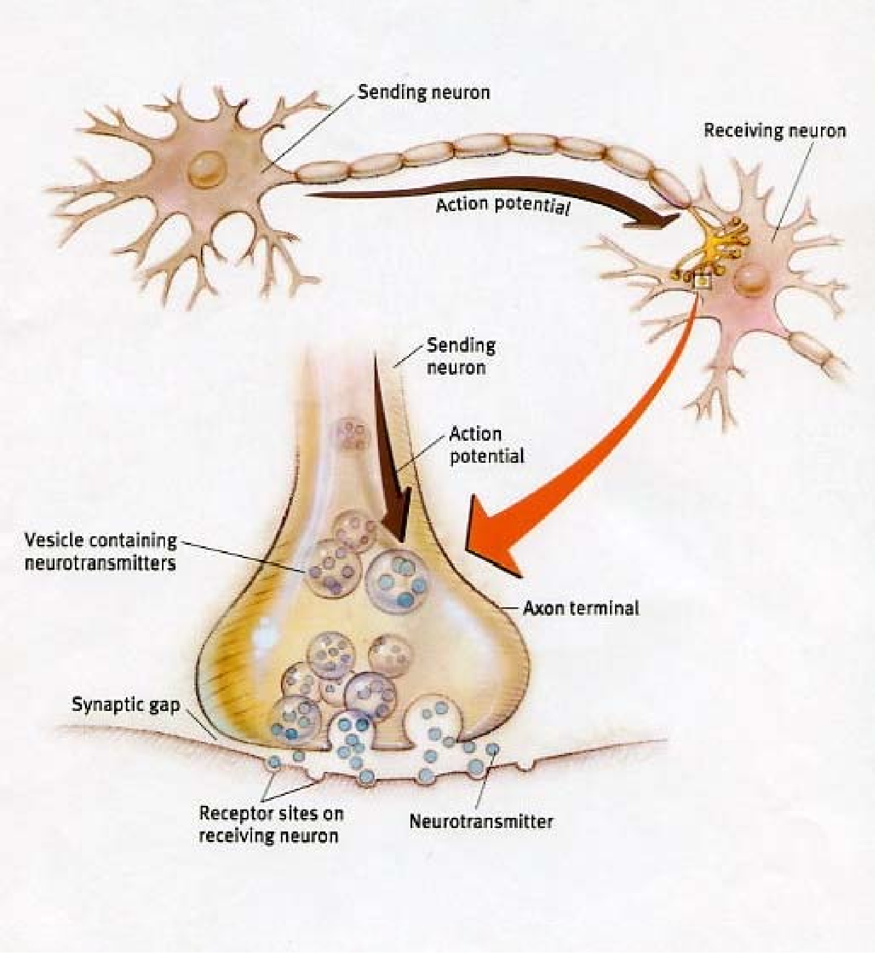
- **A feed-forward artificial network**
  - A machine learning mimicking the way **human brain processes** information when trying to make **a meaningful conclusion** based on previously seen patterns
- **The input layer**
  - A protein sequence
- **The output layer**
  - One of the **possible** outcomes
    - Whether a particular amino acid lies within an α**-helix, a** β**-strand, or an unstructured (random) region**



The flow of information:
one direction

The Neural Network

- The neural network receives its **signals** from the **input** & passes information **to the hidden layer** through a neuron, similar to how a neuron would fire across **synapse**
  - In **the hidden layer**, the **strength** or **weight** of the input being controlled by the **neurons**
    - Many or several **nodes (in input layer)** might influence the same node in the **hidden layer**
  - **No cause-&-effect rules**, only to deduce the **probable** relationship between the **input** and **output** layers
    - The only requirement is the knowledge that the input & output layers

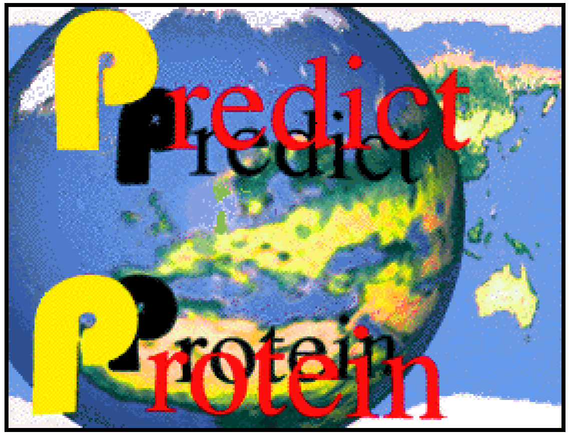


The Neural Network & PHDsec

- **A supervised learning approach** to deduce the relationship between the input and output layers
    - Based on **the training set with known answer**
        - Known three dimensional structures
            - The **secondary structure** in which a particular residue is found
            - **Other factors** influencing structural conformation

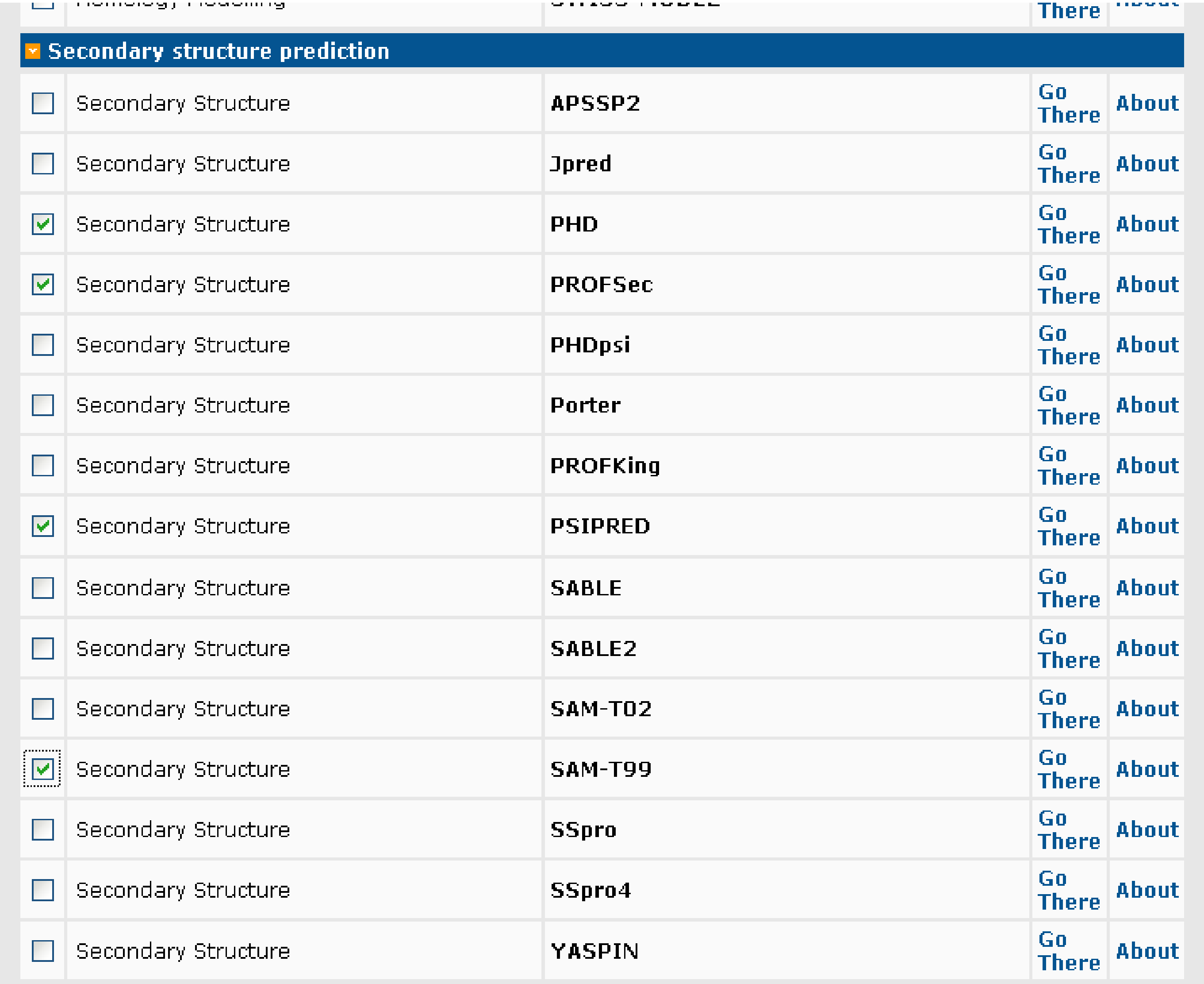
    - The network attempts to learn the relationship between **the input** & **the output layers**
        - Adjusting **the strength** of each of the interconnected neurons to fine tune the predictive power of the network

PHDsec



- PredictProtein
  1) Sequence **input**
  2) The server searches for **known homologous** proteins that are assumed **to have similar structures** as the query
     - An algorithm, called **MaxHom** (Sander & Schneider 1991) is used to produce **a multiple sequence alignment (MSA)** of all these putative homologs
       - **MaxHom** provides **a profile** of **the evolutionary history** of the sequence
         - Detailed information about **each residue**, revealing its **evolutionary conservation** & to what extent it can be **replaced** by other amino acids without changing its structure **(*****)**

- The inclusion of **evolutionary information** was **the single most important contribution** advancing the predictive power of secondary structure prediction over the last decade



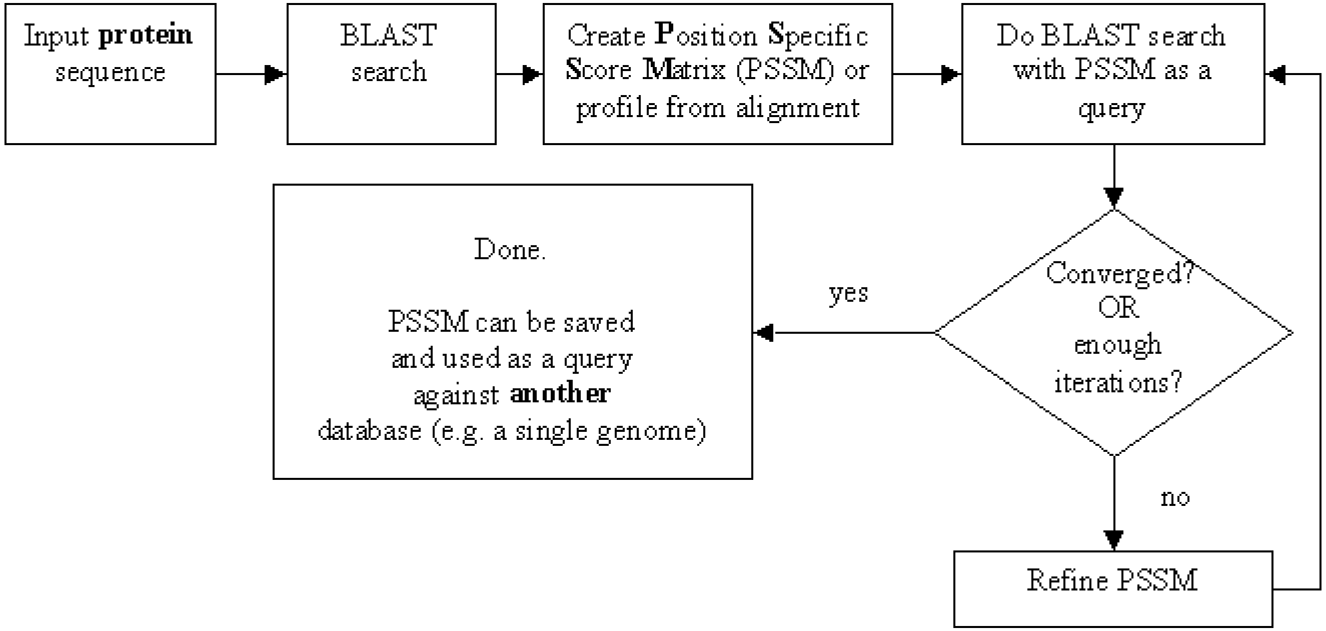
PSIPRED

- PSIPRED (McGuffin et al. 2000)
  - Based on a similar concept as **PHD**
    1) Sequence input
    2) **PSIPRED** performs **a PSI-BLAST search**
       - To compose **a profile** that conveys the evolutionary record of each position
       - The information then is fed into a system of neural network that predict secondary structure in three states

- **PSIPRED- support vector machine (SVM)-based method** (Ward et al. 2003)
  - A powerful machine-learning algorithm

- **Combined ⇨ significantly better prediction**

PSI-BLAST


Input **protein** sequence
BLAST search
Create **P**osition **S**pecific **S**core **M**atrix (PSSM) or profile from alignment
Do BLAST search with PSSM as a query
Converged? OR enough iterations?
yes
no
Refine PSSM
Done.
PSSM can be saved and used as a query against **another** database (e.g. a single genome)

© Olga Zhaxybayeva

PSI= position-specific iterated

SAM-T99

- **SAM-T99** (Karplus et al. 1998), based on a two-stage process
  1) Producing **an evolutionary profile**
     - **HMM approach**

  2) **Profile** is used as **an input** to a machine-learning program ⇨ secondary structure prediction

- **Major strength**
  - To find **remote homologs** to the query protein, i.e., proteins that are **evolutionarily related** to that sequence, but for which this relation is **difficult detect**
    - Because of their ostensible sequence divergence

HMMs in Protein Analysis (1)

- In the context of gene prediction
  - HMMs is aimed at generating **the best possible multiple sequence alignment** for a given protein family

- Considering a simple **multiple sequence alignment** of length six,

```
Q-WKPG
Q-WKPG
Q-WRPG
QIWK-G
Q-WRPG
Q-WRPG
```

- Some of the positions are absolutely **conserved**;

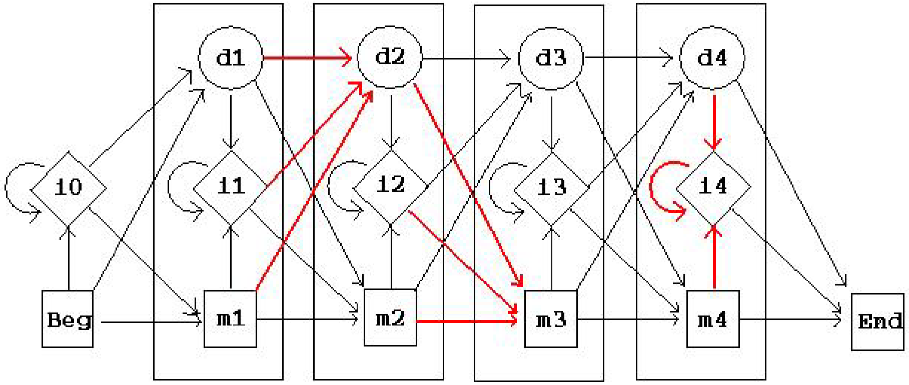
- Position 4: only by **positively charged residues**;

-

- **Gaps**: **insertion** & **deletion**

HMMs in Protein Analysis (2)

- Each of these observations can be represented by **different states in the HMMs**
- **The match state**
  - = **the most probable** amino acid(S) found at each position of the alignment
  - Match ~ misnomer
    - The match state takes into account **the probability** of finding a given amino acid **at that position of the alignment**
    - If the position is not absolutely conserve, the probabilities are adjusted accordingly, given that residues found, so
      - **Probability (match state)**
- **Insertion state**
- **Deletion state**
  - The sequence then has **to jump over a position** to continue the alignment



HMMs in Protein Analysis (3)

- The usefulness & elegance of this model comes from **the ability to train the model**
- **Without knowing** the alignment in advance
  - For **each sequence**, **the most probable path** through the model is determined
  - In turn, can be used to generate the best alignment of the sequences
- **Knowledge of these probabilities** allows for **new sequences** to be aligned to the original set or individual sequences can be **scanned against** **a series of HMMs** to see whether a new sequence of interest belongs to a previously characterized family

Evaluation of Performances (1)

- The **correct evaluation** of performance for prediction methods
  - An art itself

- **Common mistakes**
  - **A rigorous split between** the **data sets** used for **training** and **testing**
  - **Not to provide** the **standard deviations** of their **estimates**

- To compare **different methods** is very complicated
  - **Different criteria** among all methods
    - Accuracy, coverage, positive vs. negative predictive power, sensitivity, specificity…

Evaluation of Performances (2)

- Only **a handful of methods** turned out over time to not have been overestimated by their developers

- A **valid and reliable comparison** of different methods must be based on a benchmark analysis that uses
  - The same measurements, the same standards, the same sequences to all methods

- EVA (Rost & Eyrich 2001)
  - **Continuously** assesses the predictions of **automatic servers**
  - Every week, the operators of EVA receive a list of sequences of **proteins whose structures** have just been determined by experimentalists

Evaluation of Performances (3)

- EVA **automatically** submits **these sequences** as queries to all participating servers & stores the results until the release of the solved structures

- As soon as the **experimentally** solved structure is released
  - EVA assesses the predictions received from each server & score them accordingly

- The best secondary structure prediction methods
  - Now reach levels of **three-state-per-residue accuracy**
  - **Percentage** of residues predicted accurately as being in a helix, strand, or random-coil conformation **> 76%** (Rost 2001)
  - An excellent start point for secondary structure prediction

Evaluation of Performances (4)

- A comparison on results of **PROFsec** and **SAM-T99**
  - Rat cytochrome C5 fragment
  - Quite similar & similar to **experimental data**

- **Gold rule**
  - Similar results from specific methods **based on different algorithms** = the reliability is higher

- **Metaservers are better**
  - Servers that submit **the same query** to many different prediction servers & use the results to generate consensus or **jury-based metapredictions** >> single method

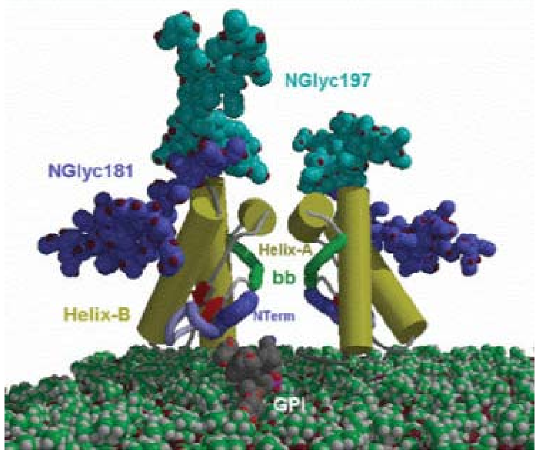
Evaluation of Performances (5)

- Different prediction methods disagree
    - Differences in the **accuracy** of different methods
    - **Genuine ambiguity** in the **definition** of secondary structure states
        - **Between** two secondary elements (end & beginning), even with experimental 3D structure resolved
            - Hard-to-define case

Evaluation of Performances (6)

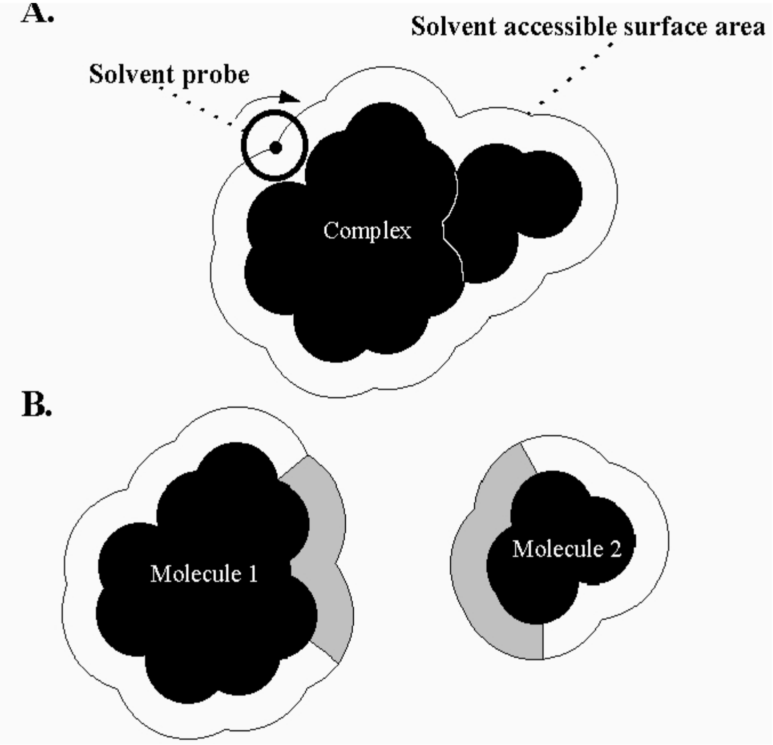
- Prion protein
  - Responsible for **aggregation** through a local flip **from helix to strand** in diseased individuals
    - Bovine spongiform encephalopathy, scrapie & Creutzfeld-Jakob disease (Prusiner 1998)
  - All methods: fail to predict as helical, instead, strand

- A computer-generated model of how a prion protein might bind to a cell membrane. This shows **two single protein molecules**, or **monomers**, bound to each other as **a double molecule, or dimer**, and anchored to the membrane surface in a vertical position
  - http://annualreport.jcs.anu.edu.au/2001/dbmb.htm



Solvent Accessibility (1)

- The area of **a protein's surface** that is exposed to the **surrounding** solvent
  - These accessible regions have the potential to **interact with** other **proteins**, **peptides**, **metal atoms** or **ions**
- Residues buried **in the interior of a protein structure** ⇨ play important role **stabilizing its structure**
  - Not be part of an active site of an enzyme, a binding site of a DNA-binding protein, or an interaction site in a signal transduction component
- **Spatial accessibility** of the residue to the solvent

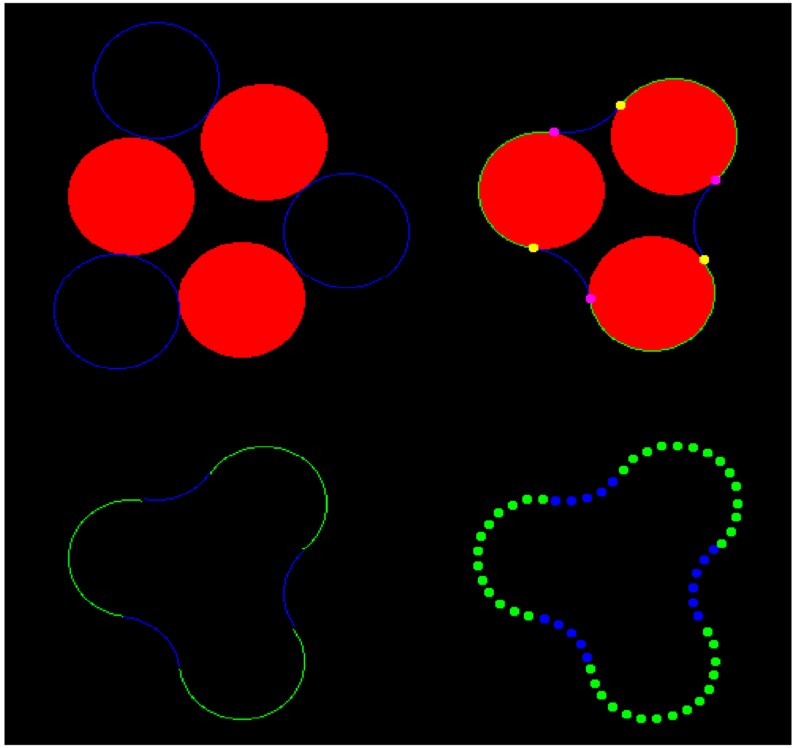


- A buried, unsatisfied hydrogen bond donor or acceptor is energetically **not good** for the folded protein

- Residues at the **surface** can be antigenic

- **Cavities (internal 'accessible' surface)** are **energetically** not good unless **'properly' filled with water**

- Accessibility calculations can also be used to determine protein-protein contact surfaces. The area where the center of the water molecule can be is called the **'accessible surface'. The surface of the molecule (black)** that can be **touched by water** is called the **'accessible molecular surface'.** The difference between these two values for the accessibile surface is very roughly a factor 3.



- There are enough reasons to think about accessible surfaces. In the figure above **the outside of the black area** is called **the molecular surface** and **the line** that indicates where the center of **the solvent probe** can be found that roles over the surface is **the accessible surface**

- The **red balls** are **atoms**. The blue balls represent **the solvent molecule** that roles over the surface. There is a little problem that is caused by the blue areas. Obviously, a water can get there, but there is no atom there to assign the accessibility to. These little blue areas are called the **reentrant surface**

Solvent Accessibility (2)

- To identify active residue to the solvent
- Most methods combine
  - Sequence profiles
  - Machine learning algorithms

Measuring Solvent Accessibility (1)

- **Square Ångstroms (Å^2)**
  - **0** (entirely buried residues) ~ **300** (residues on the surface of a protein)
  - Two entirely exposed residues may have every different **accessible areas**
    - Based simply on **the chemical structure** of the amino acid at that position

- **Assumption**: residues with **long side chains** expose to a larger area to solvent than residues with short side chain

Measuring Solvent Accessibility (2)

- The percentage of **the surface of a particular residue** that is accessible to solvents
  - Dividing **the actual accessibility** by the maximum observed for a particular amino acid
  - The amino acid type << **Ångstroms value**

- **Earlier prediction methods**
  - Thresholds that divided the 0% to 100% scale into two states
    - **Buried vs. exposed**
    - No good biophysical reason for choosing any particular **threshold**
    - 7%, 9%, 16% or 25%

Measuring Solvent Accessibility (3)

- 25%
  - ~half of all residues in a typical protein will be higher than this threshold
    - One half are exposed, the other half are buried
- http://pevsnerlab.kennedykrieger.org/bioinformatics/bioinf11_proteomics_v5.htm

PHDacc & PROFacc (1)

- Rost et al. 1996
  1. The sequence alignment & the construction of the profile
     - MaxHom
  2. The neural network
     - To assign **1 of 10 possible levels** of accessibility to each residue in the query sequence
     - The square of the returned result (the state number) indicates the % accessibility of the residue
       - 1 = 0-1%
       - 2 = 2-4%
       - Ect.

PHDacc & PROFacc (2)

- Alternative: **the 10 states** could be grouped into a two-state scheme (10-state scheme)
  - >16% of the surface area is accessible to solvent ⇨ exposed
  - ≤ 16% ⇨ buried

JPred (1)

- Guff & Barton 2000
  - A prediction service that predicts both
    - **Secondary structure** & **solvent accessibility**
  - To predict solvent accessibility
    - **HMM + PSI-BLAST**
    - A neural network uses these **profiles** to predict one of three categories of exposure
      - 0%, 5% and 25%
      - The output of predictions from **two different networks** is **combined** to give **an average relative solvent accessible**

Evaluation of Performances (1)

- Accessibility to the surrounding solvent is influenced to **a large extent** by **non-loal affects**
  - Less accuracy of prediction

- *EVA*
  - **No** such large-scale continuous system evaluates solvent accessibility

- **Two-state accessibility** predictions
  - Either **exposed or buried**
  - The accuracy of *JPred*, *PHDacc* and *PROFacc*: **75-85%**

Evaluation of Performances (2)

- **Non-binary predictions**, accuracy is more difficult to measure (not “correct” vs. “incorrect”)
  - Exposed surface **area**
  - **Percentage** of exposure

  - Possible way
    - **Correction coefficient** (CC) relating the predicted values & the actual solvent accessibility over a larger dataset
      - PHDacc: CC= **0.53** (purely random correlation = 0)
      - **Homology modeling: CC=0.66**
        - A method for **the direct prediction** of **tertiary structure**
        - Only for **very close sequence homology** with an experimentally determined structure

Evaluation of Performances (3)

- Where **known structures** **are not available**
  - Solutions: ***JPred*, *PROFacc***

- ***JPred***
  - Secondary structure prediction + solvent accessibility data

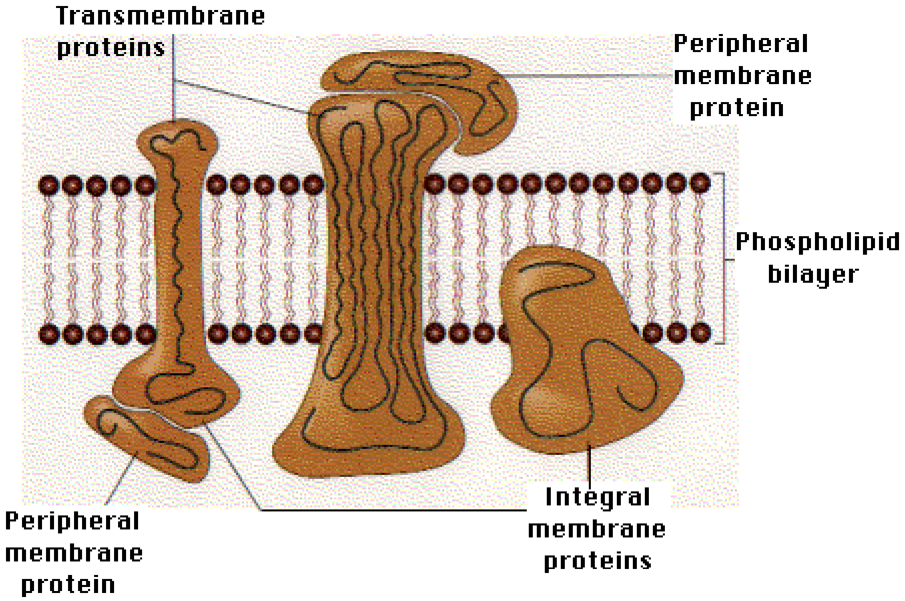
- Results **differ between methods**: **substantial**
  - Source: difference in defining exposure
    - *PHDacc*: a residue >16% accessible to solvent = exposed
    - *JPred*: **25%**

Evaluation of Performances (4)

- Example

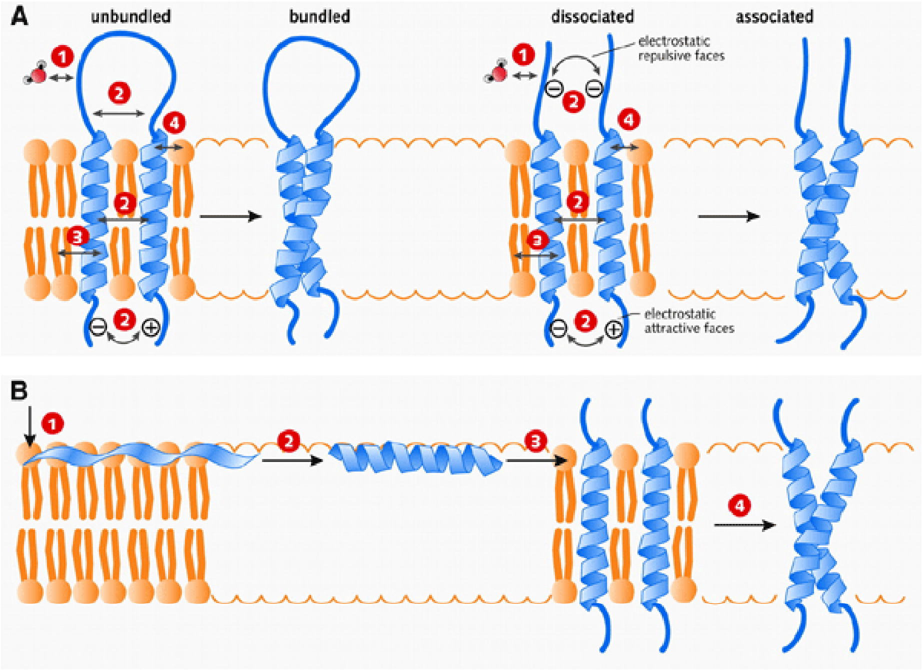
Transmembrane Segments (1)

- Proteins that are embedded in the **cell's membrane**
  - Cell interacting with other cells in its vicinity
    - Interaction with both **intracellular** and **extracellular** sides of the membranes


Transmembrane proteins
Peripheral membrane protein
Phospholipid bilayer
Peripheral membrane protein
Integral membrane proteins

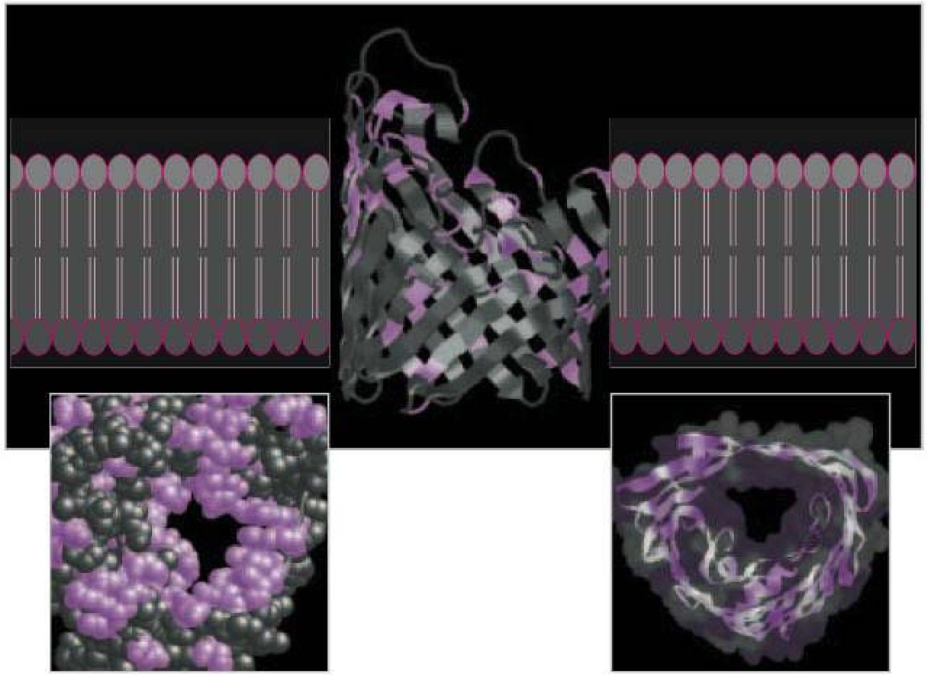
Transmembrane Segments (2)

- **~1/4 of all proteins** = membrane proteins (Melen et al. 2003)
  - Biomedicine important
  - Two broad classes
    - Insert **helices** into the lipid bilayer
    - Insert **strands**



EMBO reports **3**, 12, 1133–1138 (2002)

Transmembrane Segments (3)



Porin side view, lipophilic amino acids (gray); hydrophilic (light magenta); inset as seen from above

(PDB 3POR)

- ***Assumption***
  - The transmembrane segments of proteins share **common biophysical features**
    - The kinds of sequence that are able to **incorporate** themselves into an **environment** very **different from** that found within most cellular compartments

Transmembrane Segments (4)

- The basic biophysical requirement for **a residue** to be buried in the membrane is
  - **Hydrophobicity**
    - High degrees of **hydrophobicity** enable most **transmembrane segments** of protein actually **to remain within the membrane**, avoiding the solvent on either side

- **Prediction methods**
  - Searching for **long hydrophobic stretches of sequence**
    - Kyte & Doolittle (1982) & improvements
  - **Topology methods**
    - The orientation of membrane segments with respect to **the N-terminus** of a protein
    - Proteins that begin on the outside: **topology out**
    - Proteins that begin on the inside: **topology in**

TopPred (1)

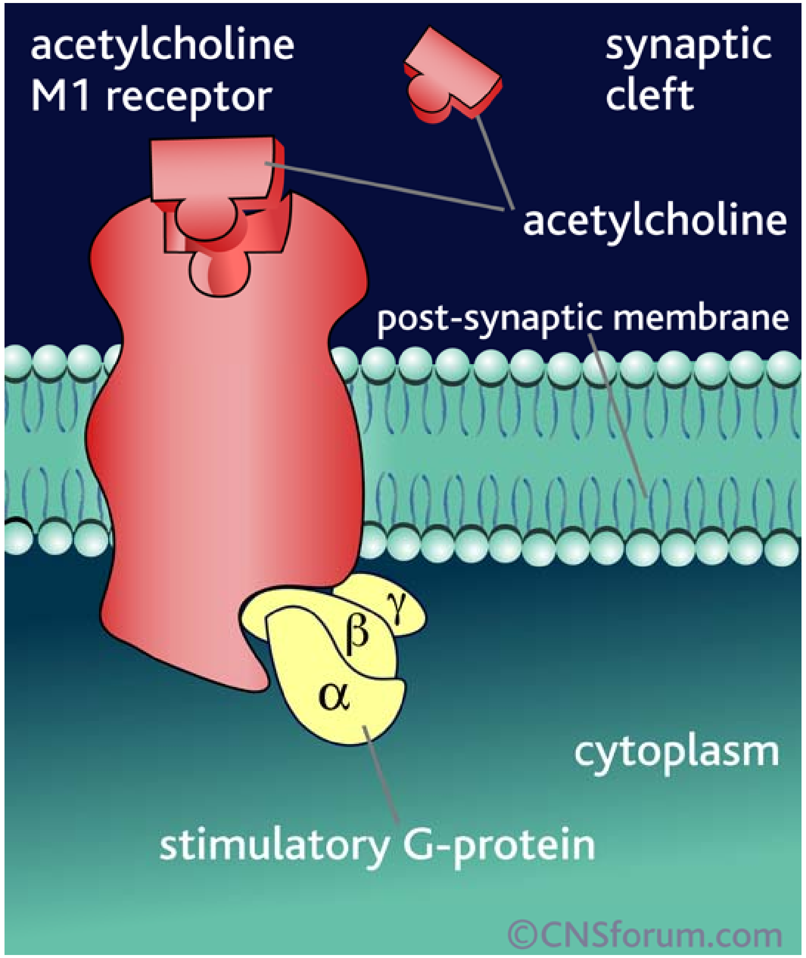
- von Heijne 1992
- **Hydrophobicity** analysis + analysis of **electrical charges**
  - The **distribution of positively charged residues** between the transmembrane helices (*****)

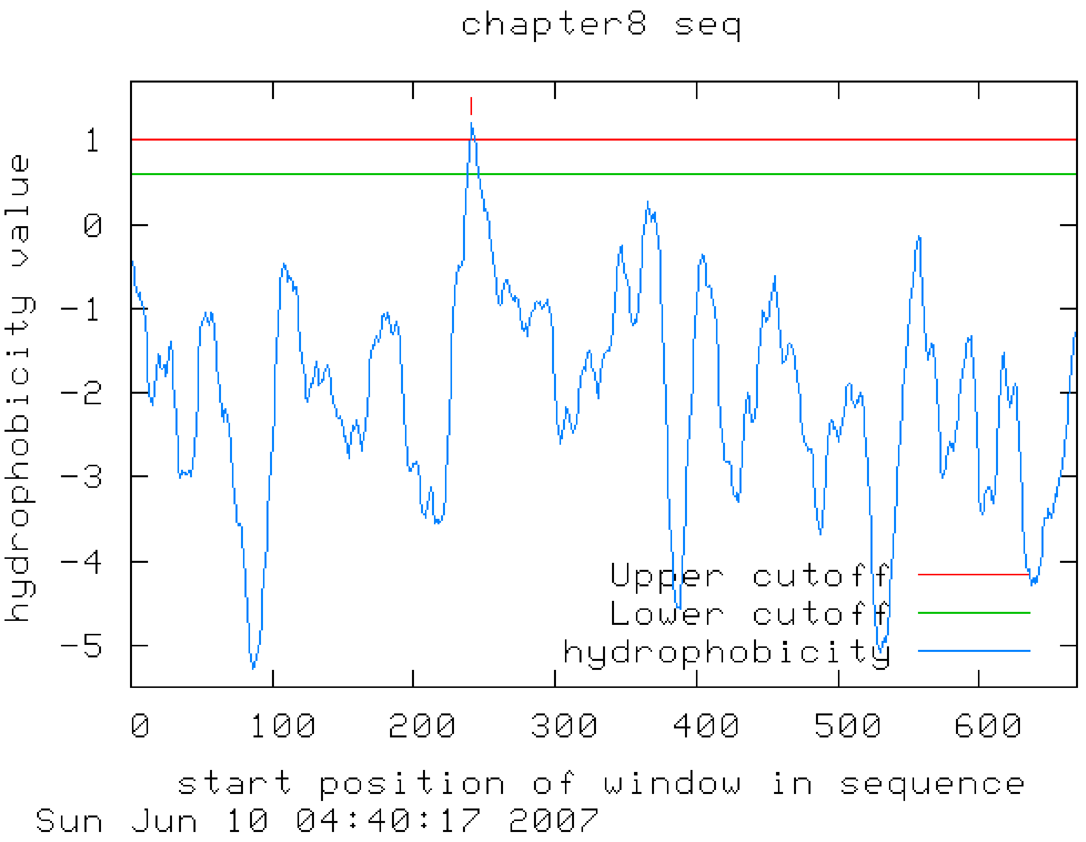
- Sequence stretches that are found to be **rich in hydrophobic** residues ⇨ **transmembrane helices**

- Stretches that are **hydrophobicity** but fail to exceed a predefined cutoff of **hydrophobicity** ⇨ **putative** transmembrane helices

TopPred (2)

- The Von Heijne positive-in rule
  - **Positively charged residues** are **more abundant** on **the inside of membrane**
- Example
  - Muscarinic acetylcholine receptor Q18007
    - M1: throughout the neurons of the central nervous system




chapter8 seq
1
0
-1
-2
-3
-4
-5
hydrophobicity value
0
100
200
300
400
500
600
start position of window in sequence
Upper cutoff
Lower cutoff
hydrophobicity
Sun Jun 10 04:40:17 2007

PHDhtm

- Rost et al. 1996; part of the PredictProtein service
  1. To construct **a profile** from **a multiple sequence alignment** (MSA)
  2. **A neural network** predicts whether each residue is likely to be part of **a transmembrane helix**
  3. ***Another neural network*** then is used to decide whether the protein as a whole is **a putative helix bundle** integral membrane protein
  4. The system predicts **the topology** of the protein by applying **the von Heijne positive-inside rule** to the network predictions

ProfTMB

- Bigelow et al. 2004

- Specializing in the prediction of **transmembrane β-strands**

- Multiple sequence alignments (MSA) to produce **a profile** that is fed into **an HMM**
  - The HMM is trained on examples from one specific group of membrane proteins **known as β-barrels**
    - Proteins that reside in the outer membrane of **gram-negative bacteria, mitochondria, and cholorplasts**

SOSUI (1)

- Hirokawa et al. 1998

- **Four main parameters**

1. To calculates the hydropathy of residues
   - Based on Kyle-Doolittle index

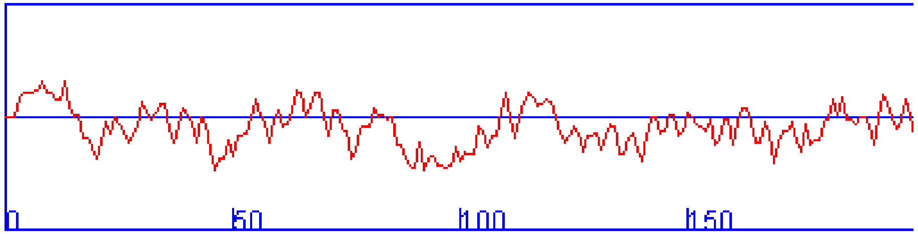
2. To calculate **the charges** of each of the residues

3. To calculate the **amphiphilicity**
   - The distribution of electric charges around the helix

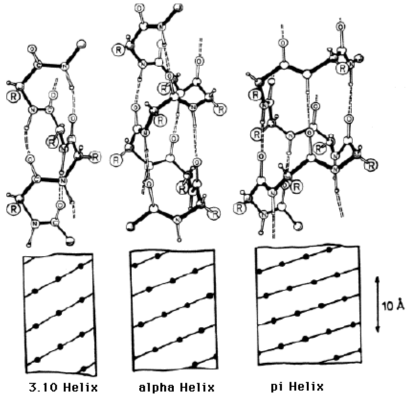
4. The length of the sequence is incorporated into the calculation

SOSUI (2)

- Hirokawa et al. 1998



- **Outputs**
  - A graph with the hydropathy profile of the query sequence
  - A “helix wheel” diagram illustrating the predicted transmembrane segments
    - The different features of the helix residues and enables visualization of the biophysical traits of the helix as a whole



TMHMM (1)

  - Krogh et al. 2001

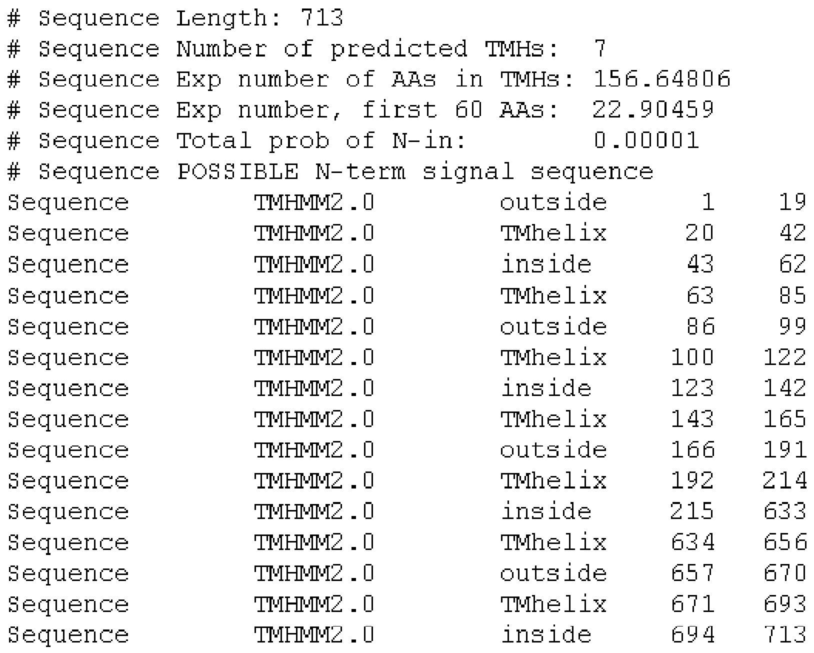
- HMMs are **particularly useful** in matching a sequence to a **predefined "grammar"**
  - Transmembrane proteins tend to obey **a relatively strict grammar** of
    - Alternating segments of membrane & non-membrane segments
    - A well-defined organization of positively charged residues

TMHMM (2)

  - Krogh et al. 2001

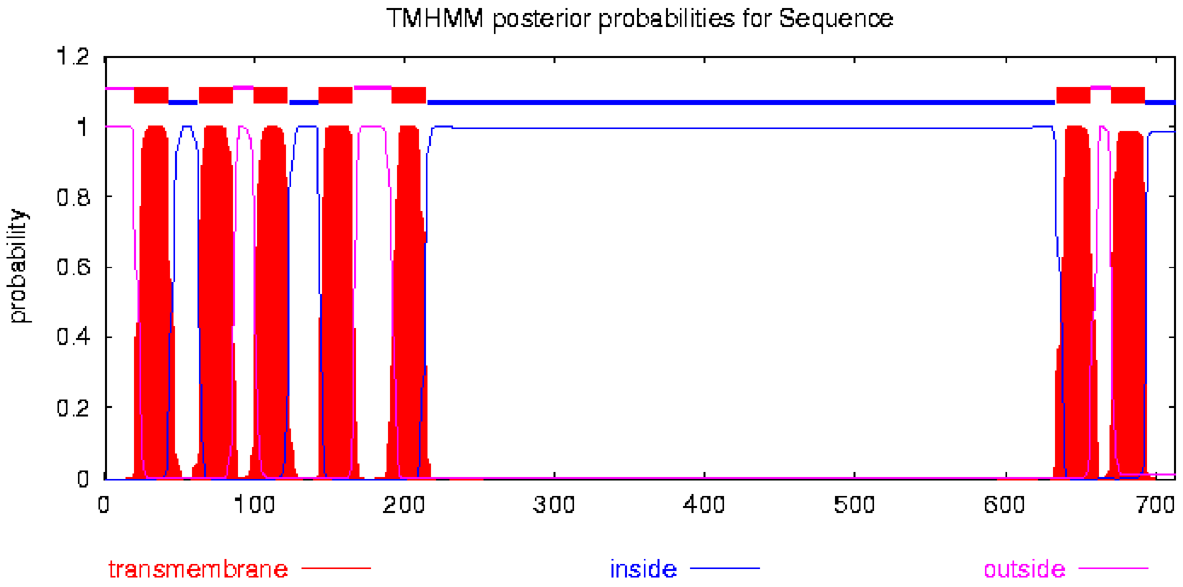
- TMHMM
  - To match the **query sequence** to this **grammar**, which is derived from a set of well-characterized transmembrane proteins
    - To predicts the segments that are most likely **to traverse the membrane** & the most likely topology of the whole protein

- Example: protein Q18007
  - Muscarinic acetylcholine receptor



```
# Sequence Length: 713
# Sequence Number of predicted TMHs:  7
# Sequence Exp number of AAs in TMHs: 156.64806
# Sequence Exp number, first 60 AAs:  22.90459
# Sequence Total prob of N-in:        0.00001
# Sequence POSSIBLE N-term signal sequence
Sequence        TMHMM2.0        outside      1     19
Sequence        TMHMM2.0        TMhelix     20     42
Sequence        TMHMM2.0        inside      43     62
Sequence        TMHMM2.0        TMhelix     63     85
Sequence        TMHMM2.0        outside     86     99
Sequence        TMHMM2.0        TMhelix    100    122
Sequence        TMHMM2.0        inside     123    142
Sequence        TMHMM2.0        TMhelix    143    165
Sequence        TMHMM2.0        outside    166    191
Sequence        TMHMM2.0        TMhelix    192    214
Sequence        TMHMM2.0        inside     215    633
Sequence        TMHMM2.0        TMhelix    634    656
Sequence        TMHMM2.0        outside    657    670
Sequence        TMHMM2.0        TMhelix    671    693
Sequence        TMHMM2.0        inside     694    713
```

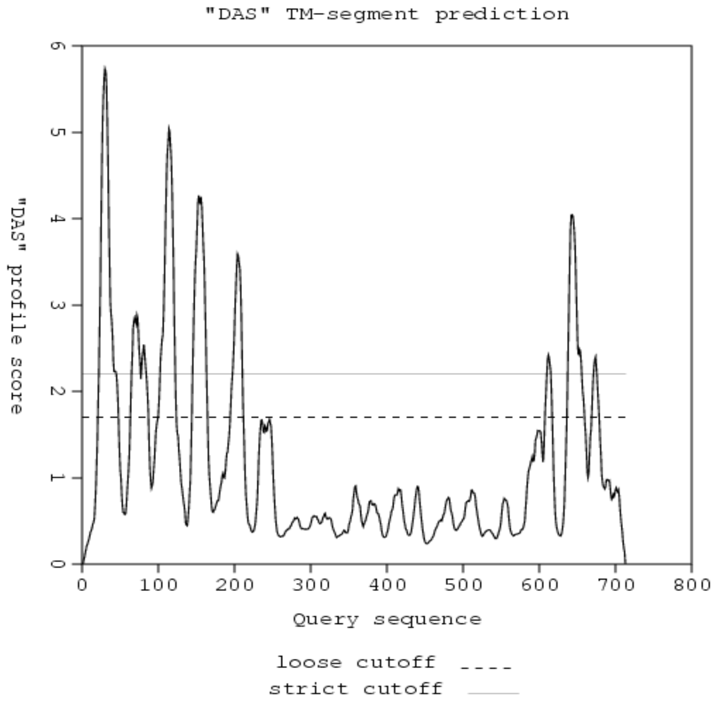
- The location of **predicted transmembrane segments**, their predicted topology, and the reliability of the prediction



DAS

- Cserzo et al. 1997

- **The dense alignment surface (DAS)**
  - To assess sequence similarities between segment s of the query protein & known transmembrane segments
    - Their biophysical similarity to stretches that were shown **experimentally** to be integrated within the membrane



Evaluation of Performances (1)

- No continuous, large-scale system to assess and compare the performances
- Developers' reports: 75-95%
  - **Overoptimistic** (Chen & Rost 2002; Chen et al. 2002)
  - Striking similarities between ***TopPred*** & ***TMHMM***
    - Q18007
    - Both methods predict that,
      - Along the first 200 residues, the protein crosses the membrane five times
      - The protein does not cross the membrane at all over the next 400 residues
      - In last 200 residues, there are two transmembrane segments

Evaluation of Performances (2)

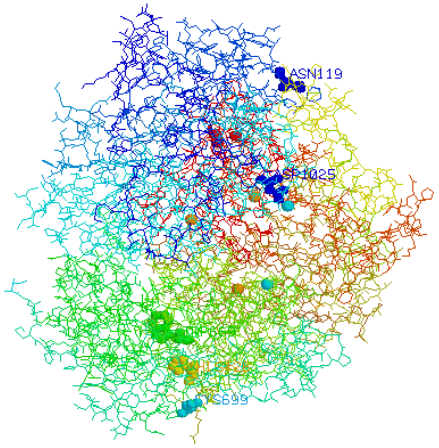
- Informative differences between ***TopPred*** & ***TMHMM***
  - On which side the long, central stretch of the protein, between positions 200 and 600, does not cross the membrane
    - Topologies???

- **Solution**
  - To submit the same sequence to yet another server, preferably one that uses **a different algorithms** than these two, and hope that the result supports one or the other, in a majority rule fashion

Predicting Function

- To extract **biologically important information** from a protein sequence
  - **Structure prediction**
    - To predict secondary structure + solvent accessibility + transmembrane helices…
  - **Function prediction**
    - Case-specific prediction
      - Did not result in automated tools for function prediction, but by
        - **Annotation transfer**
        - **Motifs & patterns**

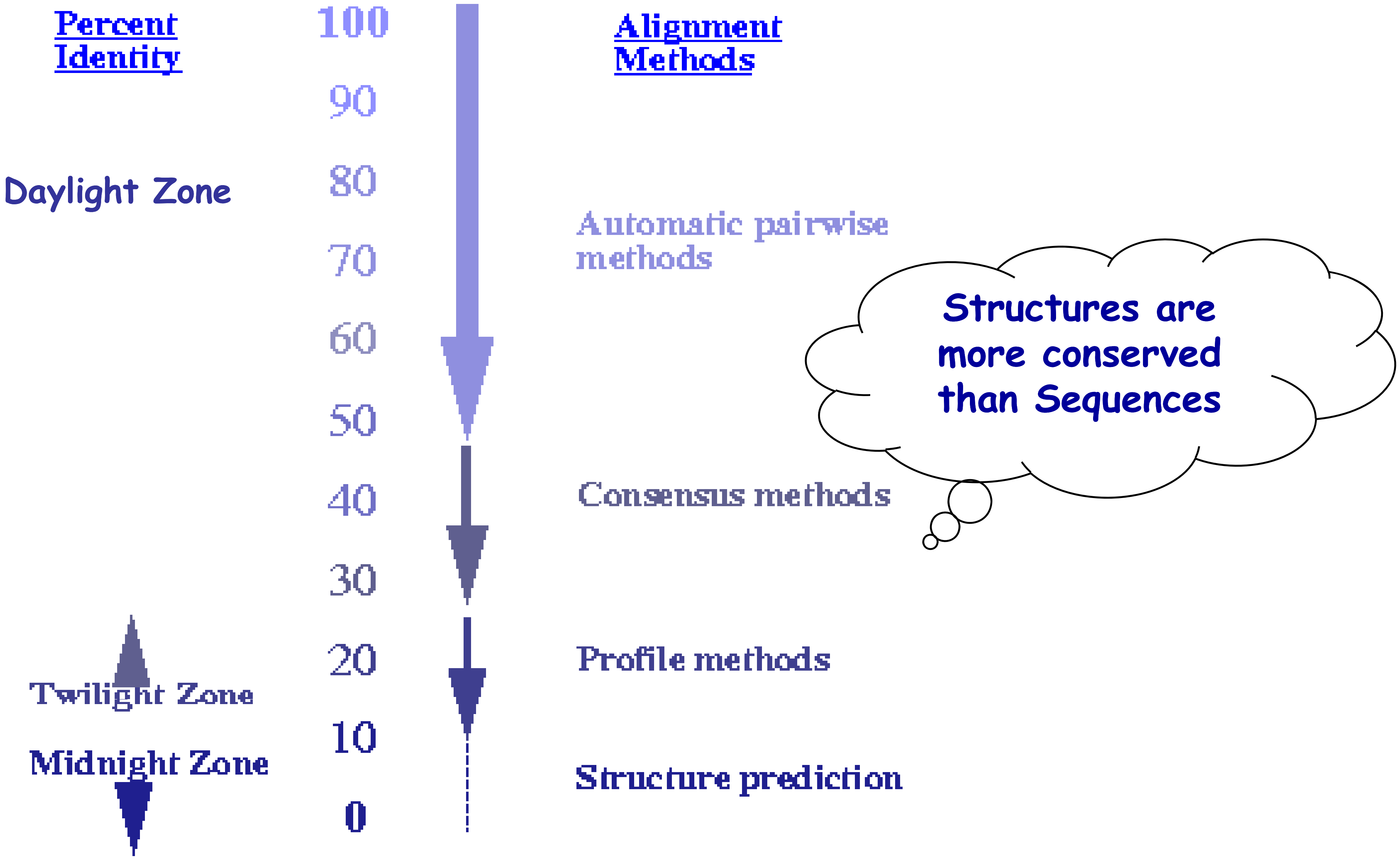
Annotation Transfer (2)



- Thousands of proteins have been characterized **experimentally** in the lab
  - Their functions have been recorded in various protein databases

  - **Similarity at the sequence** level implies **similarity of function** (Next slide)
    - If a newly discovered protein bears high similarity to another, well-characterized one, it is reasonable to assume that they have a similar function

  - Some **notable exceptions**
    - Proteins have different functions depending on their cellular location
      - **"moonlighting proteins"** (Jeffery 2003)



Percent Identity

Alignment Methods

Daylight Zone

Automatic pairwise methods

Structures are more conserved than Sequences

Consensus methods

Twilight Zone

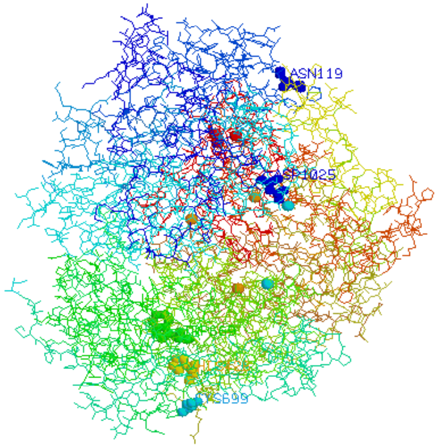
Profile methods

Midnight Zone

Structure prediction

Annotation Transfer (3)

- Automatic predictions, several important points
  - To define **a statistical threshold** of sequence similarity that permits the annotation transfer
    - This threshold differs from **one biological function to another**, needs to be determined *ad hoc* for
      - **Each protein family** &
      - **Each biological function** (Rost et al. 2003)



Motifs & Patterns (1)

- To develop various methods **to identify functionally important residues** based on their conservation throughout evolution
  - Casari et al. 1995

- Conserved **residues** & conserved **sequence elements** , it is likely to be **important** functionally
  - But no putative function yet

- The divergence between the sequence of a newly discovered protein & any other previously annotated protein is too wide to establish relatedness