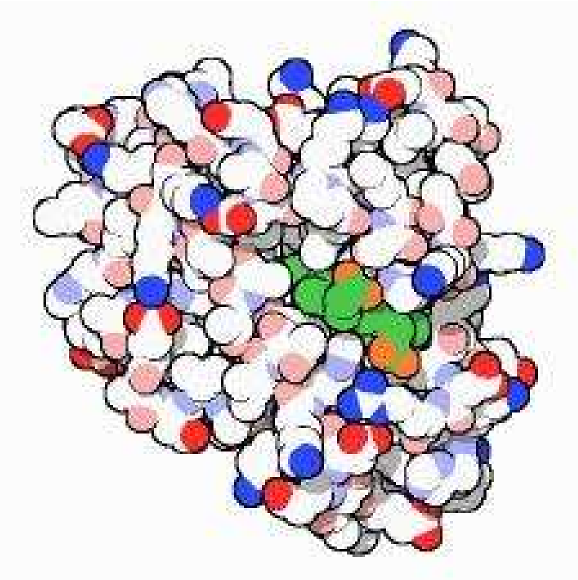
Announcements

- Homework 2 published, submit until November 30 22:00
- Journal club meetings:
  group 4 done,
  groups 2,5 met, please write a short report
  group 6 meeting tonight

Protein structure and function

Broňa Brejová

November 18, 2021

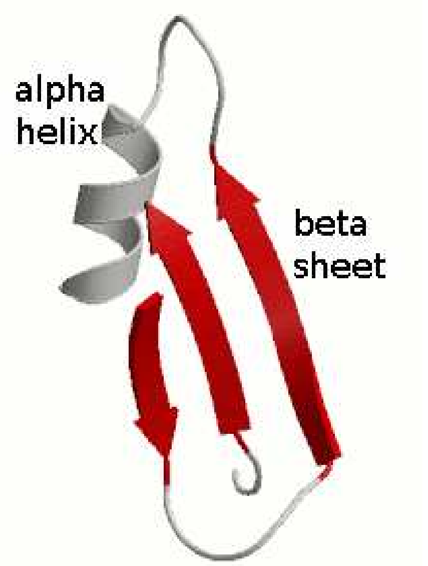


Proteins

Strings of 20 different amino acids with different chemical properties:

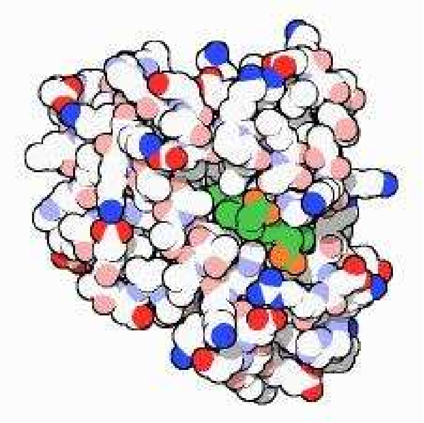
Amino Acid	Side chain	Its properties
Alanine (A)	-CH3	hydrophobic
Arginine (R)	-(CH2)3NH-C(NH)NH2	basic
Asparagine (N)	-CH2CONH2	hydrophilic
Aspartic acid (D)	-CH2COOH	acidic
Cysteine (C)	-CH2SH	hydrophobic
Glutamic acid (E)	-CH2CH2COOH	acidic
Glutamine (Q)	-CH2CH2CONH2	hydrophilic
Glycine (G)	-H	hydrophilic
Histidine (H)	-CH2-C3H3N2	basic
Isoleucine (I)	-CH(CH3)CH2CH3	hydrophobic
Leucine (L)	-CH2CH(CH3)2	hydrophobic
Lysine (K)	-(CH2)4NH2	basic
Methionine (M)	-CH2CH2SCH3	hydrophobic
Phenylalanine (F)	-CH2C6H5	hydrophobic
Proline (P)	-CH2CH2CH2-	hydrophobic
Serine (S)	-CH2OH	hydrophilic
Threonine (T)	-CH(OH)CH3	hydrophilic
Tryptophan (W)	-CH2C8H6N	hydrophobic
Tyrosine (Y)	-CH2-C6H4OH	hydrophobic
Valine (V)	-CH(CH3)2	hydrophobic

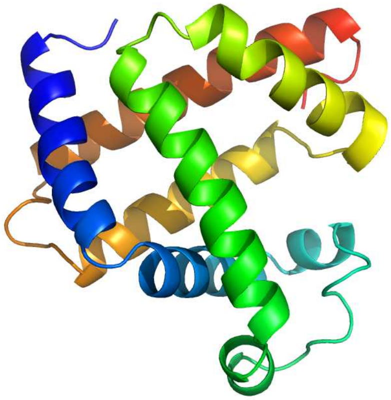
Protein structure



- **Primary structure:** sequence of amino acid
- **Secondary structure:** regular structural motifs
  alpha helix, beta sheet
- **Tertiary structure:** exact 3D positions of atoms
- **Quaternary structure:** interactions of several proteins in complex

Myoglobin, the first protein with a known structure
[Kendrew et al 1958]





Experimental structure determination

- X-ray crystallography
  – requires crystal form of the protein
- NMR (nuclear magnetic resonance spectroscopy)
  – mainly used on short proteins
- Cryo-EM (cryogenic electron microscopy)
  – less accurate, good for large protein complexes
- Expensive and difficult process
- Database of structures PDB
  184 000 protein structures
  (UniProt has over 200 million of sequences)

Bioinformatics problem: protein structure prediction, protein folding

Input: protein sequence

Output: 3D positions of atoms or amino acids

Ab initio methods

- Find a structure with the lowest free energy
- Physics-based formulas for approximating energy
  – forces among atoms of the protein and surrounding water
- Very hard computational problem
  – molecular dynamics simulation
  – optimization methods, e.g. gradient descent, simulated annealing
- Useful for short proteins and improving approximate structures

Practical approaches to protein structure prediction

For a **query protein**:

- Check if it has a **known structure** in PDB
- If not, try to find a **similar protein** in PDB (BLAST), query likely a similar structure
- If no appropriate BLAST match, try to find similar proteins by more sensitive approaches, **protein profiles** (this lecture)
- Even more distant homology can be found by **protein threading**
- Recently, approaches based on **deep learning** (neural networks) quite successful
- We can try to improve found structures by **energy minimization**
- **Predicted structures** can be also found in databases

Protein threading

- Even proteins with very different sequences can have similar structures
- We can try to “thread” the query protein to each known structure
- A special form of alignment taking into account interactions of amino acids in the known structure
- Computationally hard problem

Newest approaches: deep neural networks

- CASP competition every two years
- In 2018, 2020 won by AlphaFold designed by DeepMind/Google.
  In 2020, AlphaFold won by a large margin,
  predicted very well 2/3 of structures.
  It combines new ideas and existing approaches.
- Key idea used already before AlphaFold: **co-evolution detection**
  Find many homologs of the query protein
  (even if no structure known),
  build a multiple alignment,
  find positions that change together in evolution,
  these are potential 3D contacts

Newest approaches: deep neural networks

- **AlphaFold 1 (2018):**
  (1) Prediction of amino acid distances by a neural network.
  (2) Finding structure agreeing well with distances
  and an energy model using standard numerical optimization
  (gradient method) [animation]

- **AlphaFold 2 (2020):**
  combines both steps to a single neural network,
  which is run repeatedly on its outputs

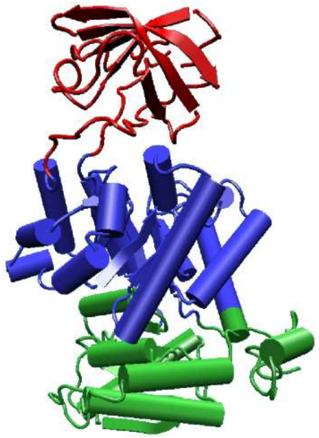
Recall: Practical approaches to protein structure prediction

For a **query protein**:

- Check if it has a **known structure** in PDB
- If not, try to find a **similar protein** in PDB (BLAST), query likely a similar structure
- If no appropriate BLAST match, try to find similar proteins by more sensitive approaches, **protein profiles** (this lecture)
- Even more distant homology can be found by **protein threading**
- Recently, approaches based on **deep learning** (neural networks) quite successful
- We can try to improve found structures by **energy minimization**
- **Predicted structures** can be also found in databases

Protein domains and families

Domain (doména)



- Part of a protein with an independent structure
- Many proteins contain multiple domains
- Domains can be rearranged during evolution

Family (rodina)

- Group of proteins or domains with similar sequence, structure and function
- If we know the structure of one family member, others might have a similar structure

Proteins as mosaics of domains

Pfam database

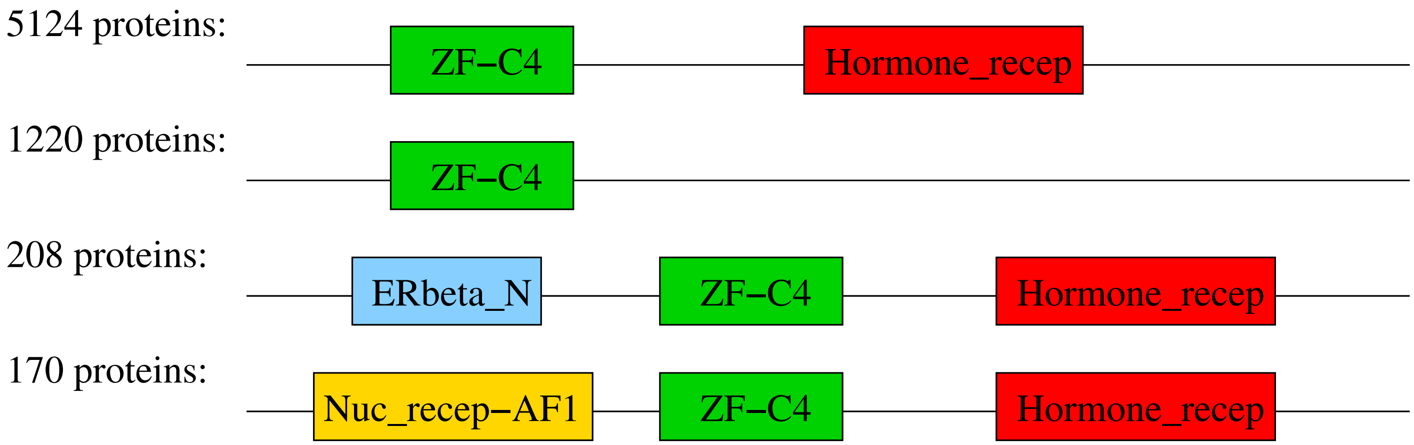
Domains in proteins classified to over 18 thousand families

77% of proteins have at least one known domain

53% protein sequences are covered by known domains

Example:

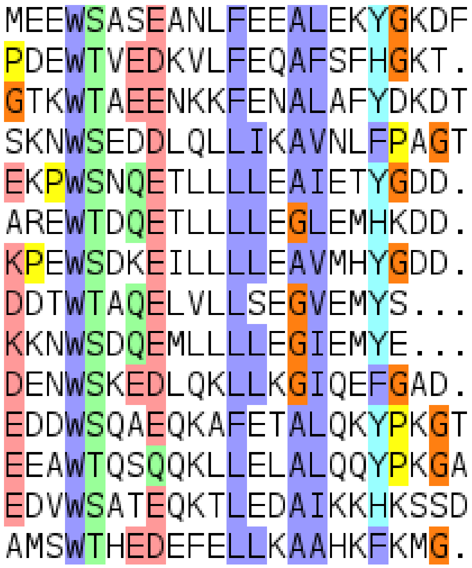
4 out of 91 architectures with Zinc finger, C4 type domain (Pfam)



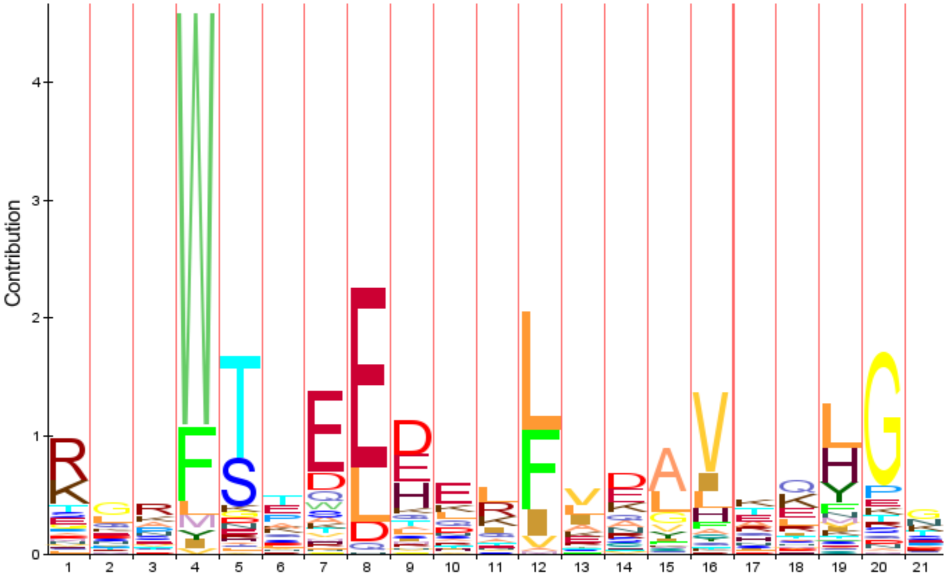
Characterization of a protein family

- Pairwise alignments (BLAST) between a query protein and family members do not always find weaker similarity

- Multiple sequence alignment of a family highlights important conserved positions



```
MEEWSASEANLFEEALEKYGKDF
PDEWTVEDKVLFEQAFSFHGKT.
GTKWTAEENKKFENALAFYDKDT
SKNWSEDDLQLLIKAVNLFPAGT
EKPWSNQETLLLLEAIETYGDD.
AREWTDQETLLLLEGLEMHKDD.
KPEWSDKEILLLLEAVMHYGDD.
DDTWTAQELVLLSEGVEMYS...
KKNWSDQEMLLLLEGIEMYE...
DENWSKEDLQKLLKGIQEFGAD.
EDDWSQAEQKAFETALQKYPKGT
EEAWTQSQQKLLELALQQYPKGA
EDVWSATEQKTLEDAIKKHKSSD
AMSWTHEDEFELLKAAHKFKMG.
```



Probabilistic profile of a family

(profile, position specific score matrix PSSM)

- In an alignment, compute $e_i(x)$: frequency of amino acid x in column i
- Create a model which generates sequence $x_1, x_2, \ldots, x_n$ with probability
$$e_1(x_1) \cdot e_2(x_2) \cdots e_n(x_n)$$
- Background model: sequence was generated randomly with amino acid x having frequency $q(x)$
- Score: log likelihood ratio in the two models

$$\log \frac{\prod_{i=1}^{n} e_i(x_i)}{\prod_{i=1}^{n} q(x_i)} = \sum_{i=1}^{n} \log \frac{e_i(x_i)}{q(x_i)} = \sum_{i=1}^{n} s_i(x_i)$$

Toy example of an PSSM

- Consider only leucine L a alanine A
- Multiple alignment of 10 sequences has the following counts:

	1	2	3	4
A	2	6	9	1
L	8	4	1	9

- Background model $q(A) = 30\%$, $q(L) = 70\%$
- Probability of sequence LAAL
  – in the profile model: $0.8 \cdot 0.6 \cdot 0.9 \cdot 0.9 = 0.3888$,
  – in the background model: $0.7 \cdot 0.3 \cdot 0.3 \cdot 0.7 = 0.0441$
- Score for LAAL: $\log_2(0.3888/0.0441) = 3.14$
- Score for LALA: $\log_2(0.0048/0.0441) = -3.20$

Toy example of an PSSM

- Multiple alignment of 10 sequences has the following counts:

	1	2	3	4
A	2	6	9	1
L	8	4	1	9

- Background model $q(A) = 30\%$, $q(L) = 70\%$
- Score of alanine in column 1: $s_1(A) = \log_2(0.2/0.3) = -0.58$, score of leucine in column 1: $s_1(L) = \log_2(0.8/0.7) = 0.19$
- Entire score table:

	1	2	3	4
A	-0.58	1.00	1.58	-1.58
L	0.19	-0.81	-2.81	0.36

- Score of LAAL is $0.19 + 1 + 1.58 + 0.36 = 3.13$
  Score of LALA is $0.19 + 1 - 2.81 - 1.58 = -3.20$

Pseudocounts

If some amino acid is completely absent at a given position, it would get probability 0 in the model

	1	2	3	4
A	2	6	9	0
L	8	4	1	10

To avoid this problem, add a small value, pseudocunt, to each count in the table (e.g. add 0.5):

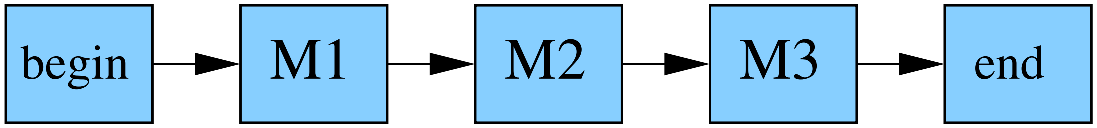
	1	2	3	4
A	2.5	6.5	9.5	0.5
L	8.5	4.5	1.5	10.5

Then compute scores as before

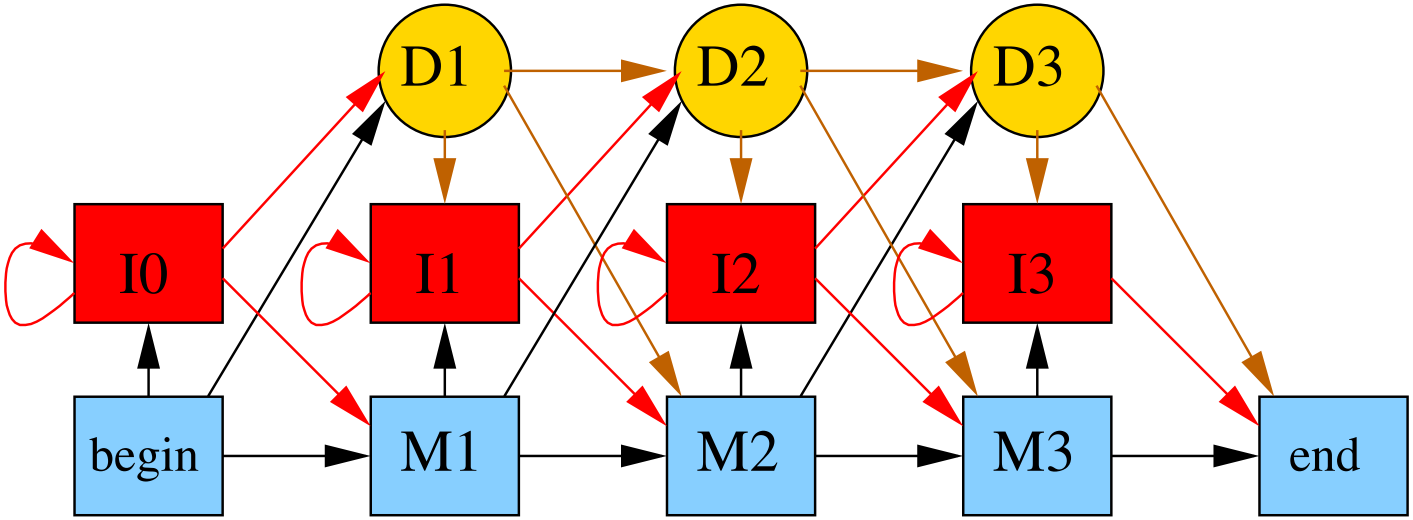
Profile HMMs (profilové HMM)

Extend profiles with insertions and deletions

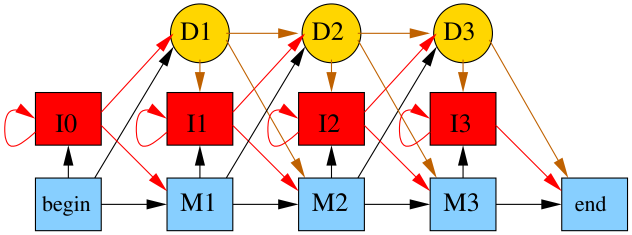
PSSM as an HMM:



Profile HMM: match state, insert state, delete state



Constructing profile HMMs



- Start from a multiple alignment
- Columns with a small fraction of gaps converted to match states, remaining columns handled by insert states
- In each column compute $E_i(a)$: the number of occurrences of a
- Emission probability $e_i(a) = \frac{E_i(a)}{\sum_b E_i(b)}$
- We add pseudocounts to avoid zero probabilities, $e_i(a) = \frac{E_i(a)+c}{\sum_b (E_i(b)+c)}$
- Transition probabilities set according to gaps
- Groups of very similar sequences used with lower weights

Using profiles and profile HMMs

Where to get profiles / profile HMMs?

- Pfam database contains domain families represented as profile HMMs
- PSI-Blast creates PSSMs on the fly from similar proteins
- PSSMs are also used to present binding site motifs in DNA (lecture on regulation)

How to find profile occurrences in a protein sequence?

- Similar to local alignemnt
- PSSM profiles: dynamic programming with fixed gap scores
- Profile HMMs: Viterbi/forward algorithms

Use the resulting score / probability to decide if a protein belongs to the family

Recall: Practical approaches to protein structure prediction

For a **query protein**:

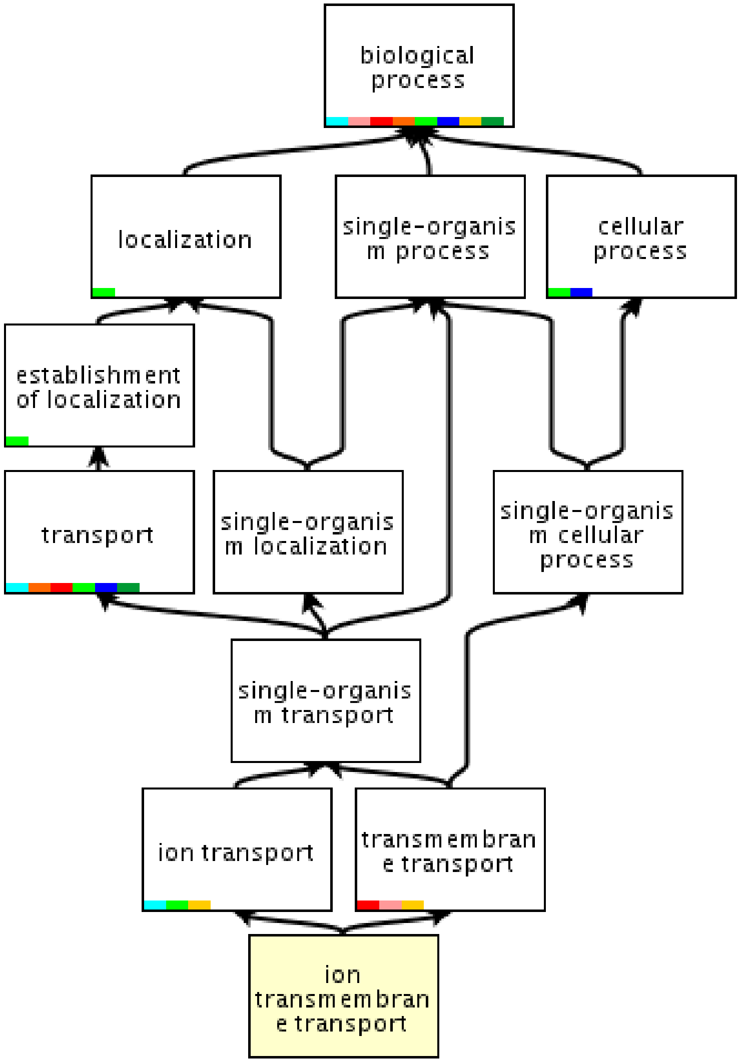
- Check if it has a **known structure** in PDB
- If not, try to find a **similar protein** in PDB (BLAST), our query likely has a similar structure
- If no appropriate BLAST match, try to find similar proteins by more sensitive approaches, **protein profiles** (this lecture)
- Even more distant homology can be found by **protein threading**
- Recently, approaches based on **deep learning** (neural networks) quite successful
- We can try to improve found structures by **energy minimization**
- **Predicted structures** can be also found in databases

Protein function

- Determined experimentally for some proteins
- Transfered to other proteins based on sequence similarity, domains, position in the genome and other data
- Swissprot/Uniprot collects known information about protein function
- Protein classification using Gene ontology (GO)
  Example of a term in GO:
  Accession: GO:0034220
  Name: ion transmembrane transport
  Ontology: biological_process
  Definition: A process in which an ion is transported from one side of a membrane to the other by means of some agent such as a transporter or pore.
  Comment: Note that this term is not intended for use in annotating lateral movement within membranes.

Gene ontology (GO)

Hierarchy of terms:

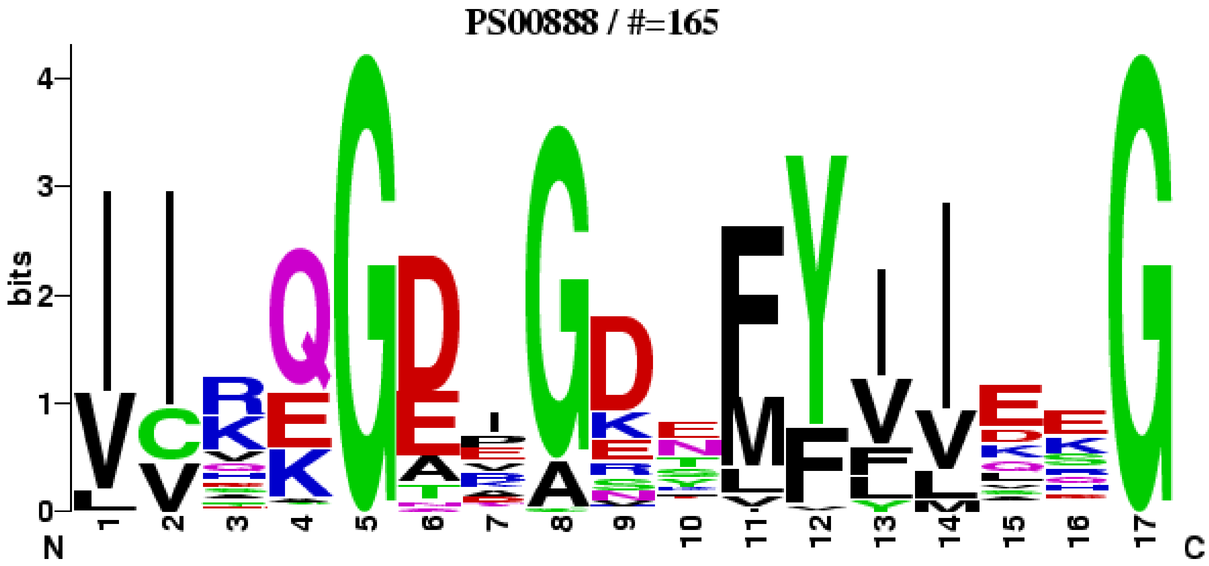

biological process
localization
single-organism process
cellular process
establishment of localization
transport
single-organism localization
single-organism cellular process
single-organism transport
ion transport
transmembrane transport
ion transmembrane transport

Other examples of HMM and profile use in protein analysis

- Predicting secondary structure
- Predicting transmembrane proteins and signal peptides
- Predicting functional motifs and posttranslational modifications (PROSITE database)

Cyclic nucleotide-binding domain signature 1:

`[LIVM]-[VIC]-x-{H}-G-[DENQTA]-x-[GAC]-{L}-x-[LIVMFY](4)-x(2)-G`


PS00888 / #=165
bits
N
C