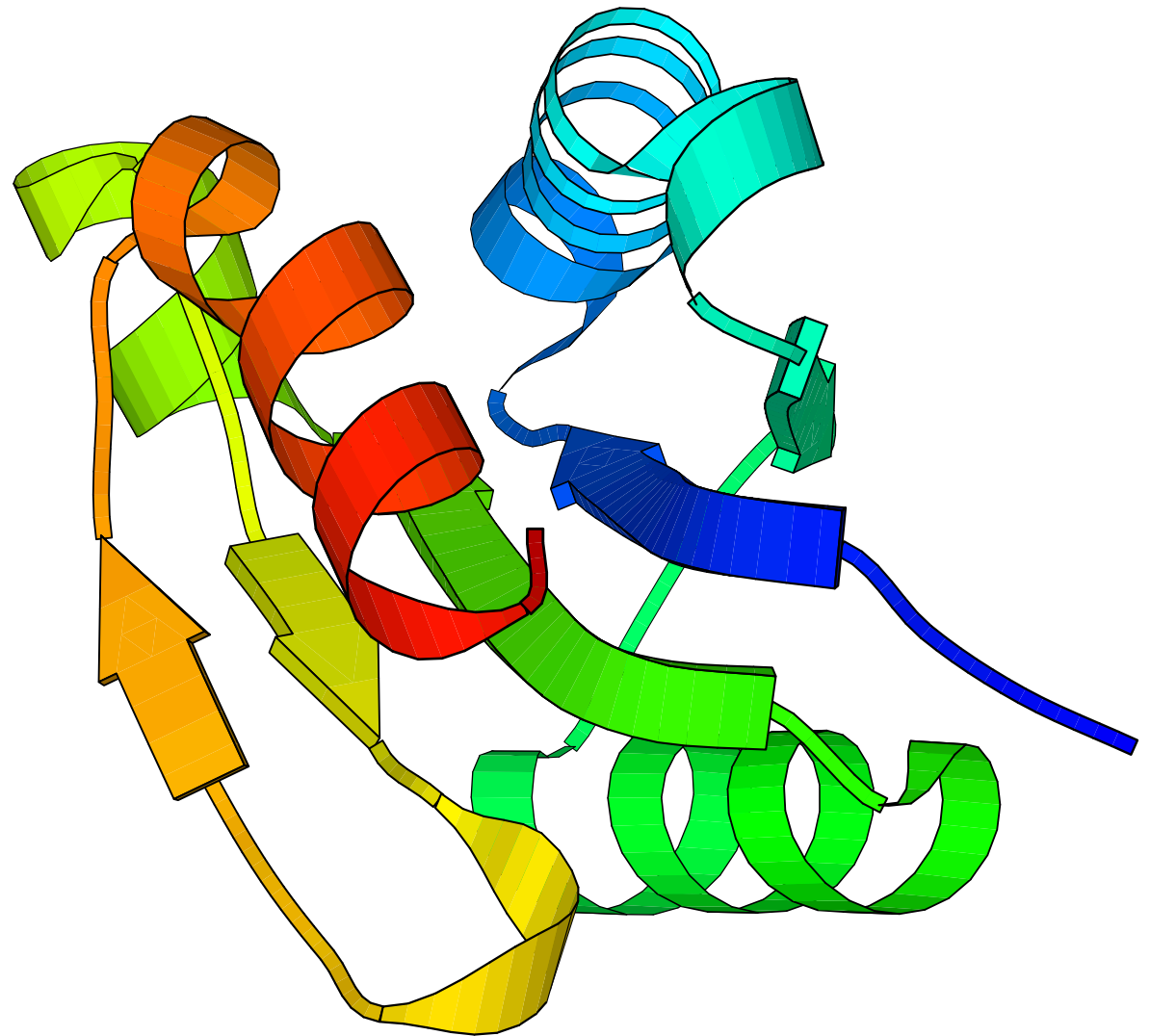


Protein Structure Prediction I

Bob MacCallum, Stockholm Bioinformatics Center

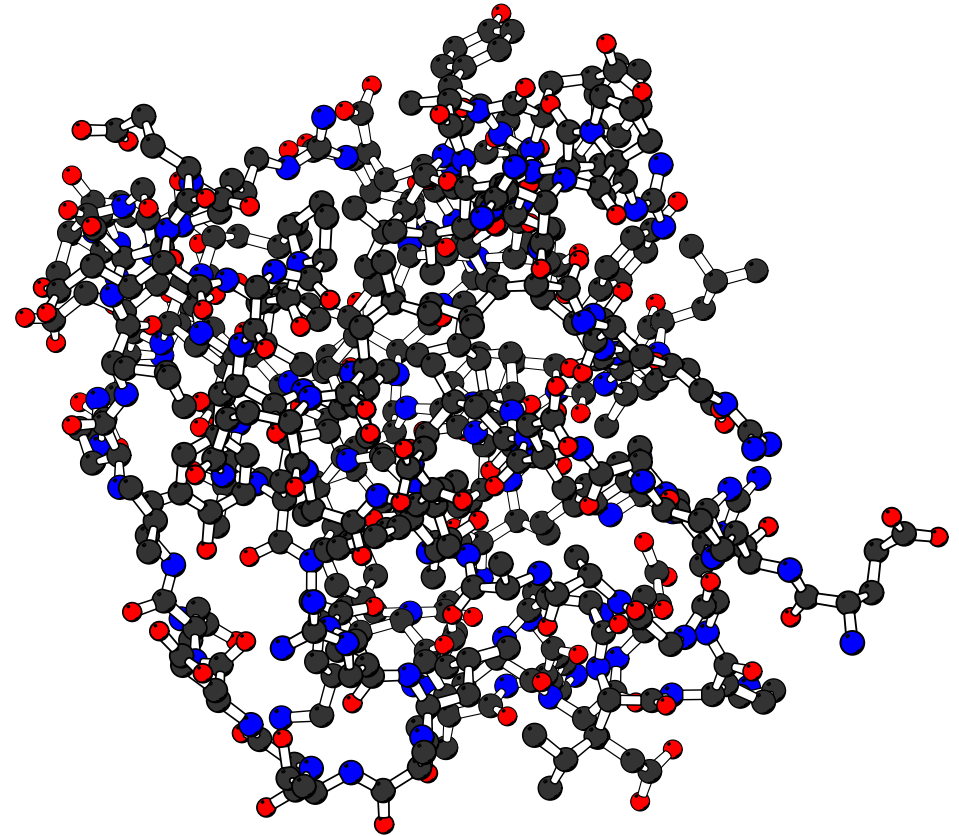
```
EKGPDLYLIPLTEEAVAEAF  
YLAEALRPRLRAEYALAPRK  
PAKGLEEALKRGAAGFLG  
EDEL RAGEVTLKRLATGEQV  
RLSREEVPGYLLQALG
```

+ computer =



It's not that simple...

Amino acid sequence contains all the information for 3D structure
(experiments of Anfinsen, 1970's)



But, there are thousands of atoms, rotatable bonds, solvent and other molecules to deal with...

Why do we need structure prediction?

3D structure give clues to function:

- active sites, binding sites, conformational changes...
- structure and function conserved more than sequence

3D structure determination is difficult, slow and expensive

Intellectual challenge, Nobel prizes etc...

Engineering new proteins

Structure prediction

Summary of the four main approaches to structure prediction. Note that there are overlaps between nearly all categories.

Method	Knowledge	Approach	Difficulty	Usefulness
Comparative modelling (Homology modelling)	Proteins of known structure	Identify related structure with sequence methods, copy 3D coords and modify where necessary	Relatively easy	Very, if sequence identity > 40% → drug design
Fold recognition	Proteins of known structure	Same as above, but use more sophisticated methods to find related structure	Medium	Limited due to poor models
Secondary structure prediction	Sequence-structure statistics	Forget 3D arrangement and predict where the helices/strands are	Medium	Can improve alignments, fold recognition, <i>ab initio</i>
<i>ab initio</i> tertiary structure prediction	Energy functions, statistics	Simulate folding, or generate lots of structures and try to pick the correct one	Very hard	Not really

CASP

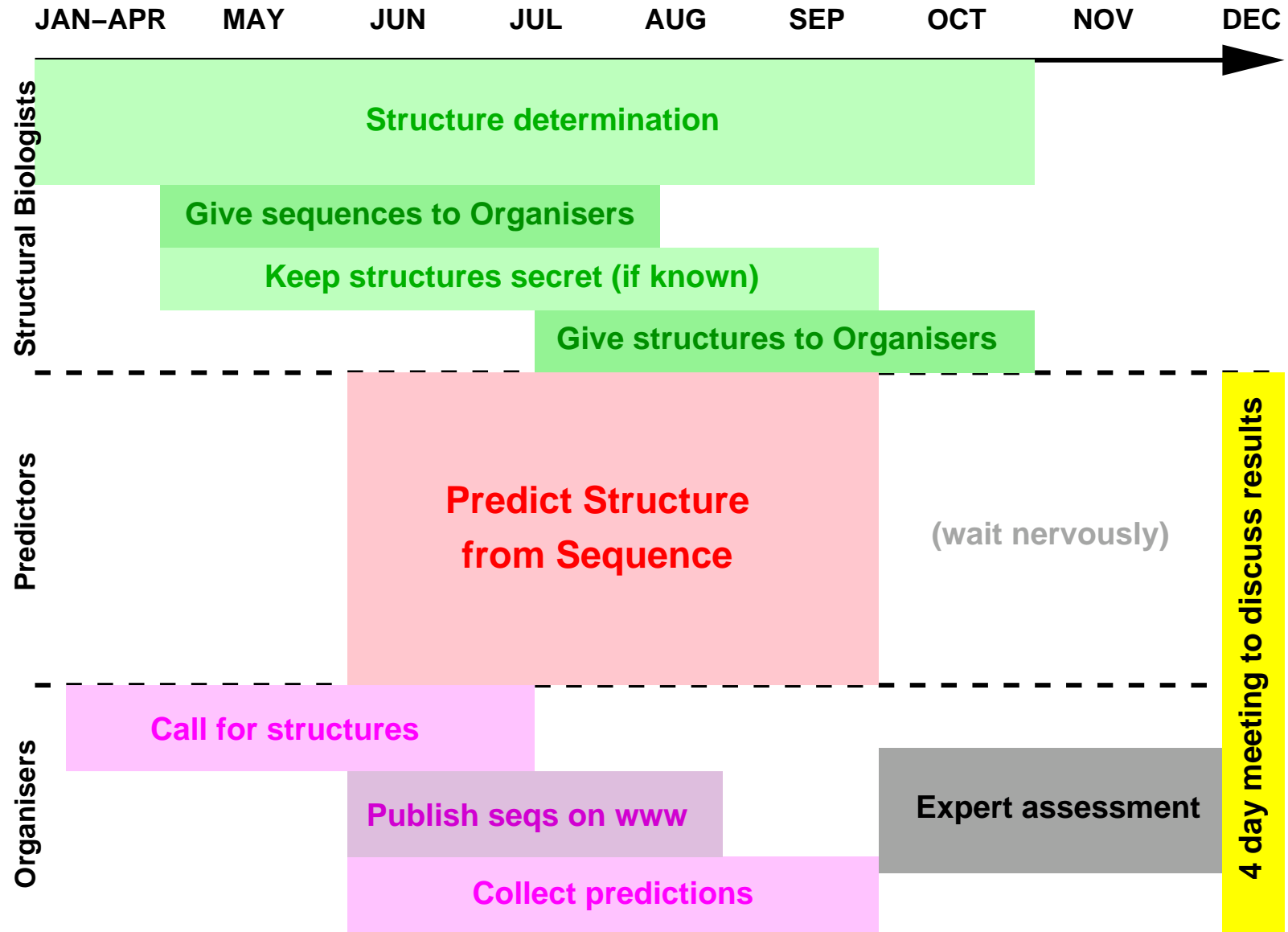
Critical Assessment of Techniques for Protein Structure Prediction

Why do we have CASP?

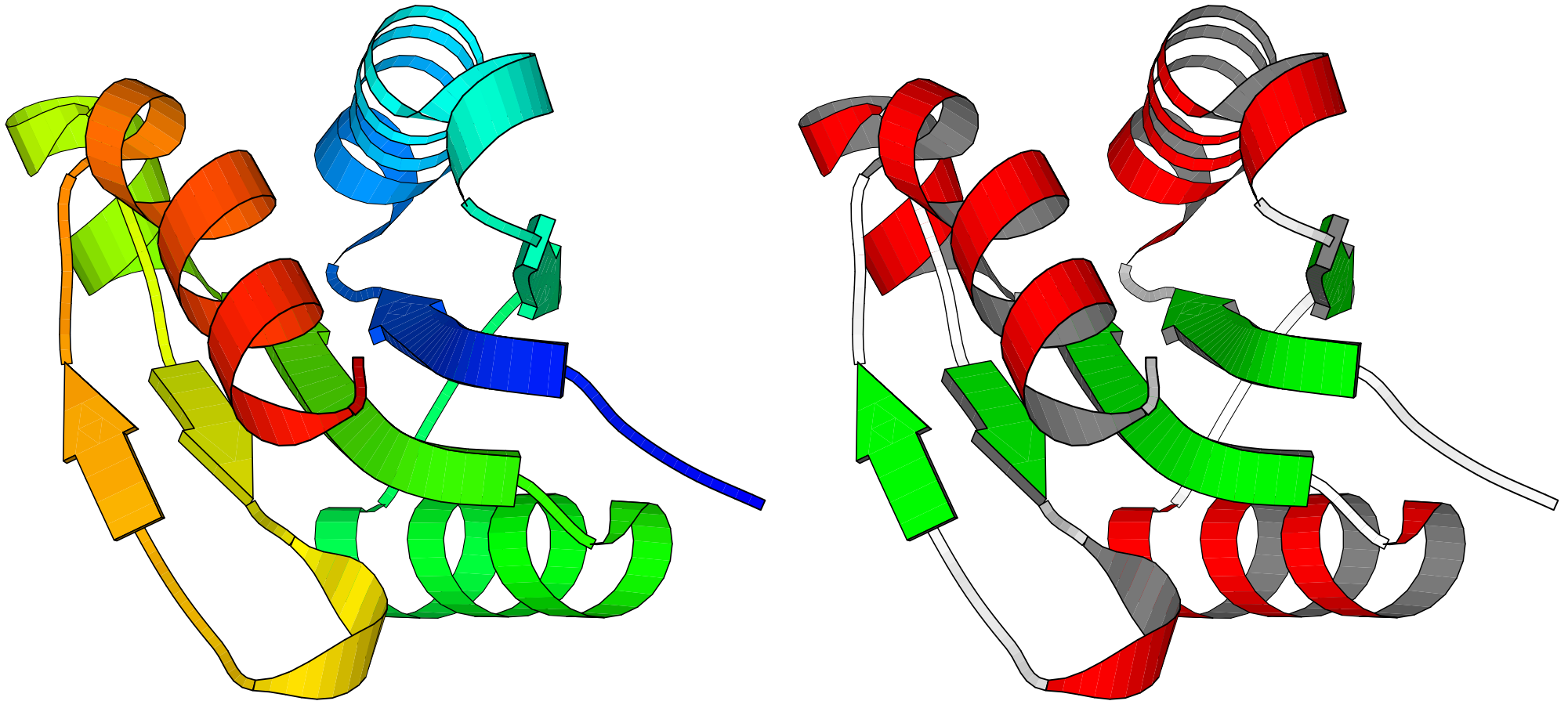
- People cheat!
 - people work hard to make prediction programs work for their favourite proteins, but...
 - benchmarking may be polluted by “information leakage”
- Difficult to compare methods fairly
 - software and data issues
 - different measures, standards

What we want is fully blind trials of prediction methods by a third party.

CASP



Secondary structure prediction (SSP)



EKGPDLYLIPLTEEAVAEAFYLAELRPRLRAEYALAPRKPAKGLLEEALKRGAAGFAGFLGEDEL RAGEVTLKRLATGEQVRLSREEVPGYLLQALG
CCCC~~EEEE~~C~~HHHHHHHHHHHHHH~~CCCC~~EE~~CCCC~~HHHHHHHHHH~~CCCC~~EEEE~~C~~HHHHHH~~~~EEEEEE~~CCCC~~EEEE~~C~~HHHHHHHHHH~~C

H = Helix, E = strand (Extended conformation), C = Coil (or loop or nothing)

Secondary structure prediction

Ignore 3D, it's too hard!

Usually concentrate on helix, strand and “coil” .

Pattern recognition, but which patterns?

- some amino acids have preferences for helix or strand; due to geometry and hydrogen bonding
- spatial (along sequence) patterns, alternating hydrophobics (helical wheel)
- conservation (down alignment) in different members of protein family; insertions and deletions

Three main generations/stages in SSP method development since 1970's.

Secondary structure prediction

1st generation methods

1973-1974: only a few 3D structures existed!

Chou & Fasman classified the amino acids (from observations in 15 proteins)

	Helix	Strand
Strong former	E A L	M V I
Former	H M Q W V F	C Y F Q L T W
Weak former	K I	A
Indifferent	D T S R C	R G D
Breaker	N Y	K S H N P
Strong breaker	P G	E

Then made up some rules for helix/strand “nucleation” and “extension”. A few amino acids have special meanings at a particular end of helix/strand.

Even more complex rules (Lim, 1974), and neighbour information (GOR method) was also used

Claims of around 70-80% - actual accuracy about 50-60%

Secondary structure prediction

2nd generation methods

- sequence-to-structure relationship modelled using more complex statistics, e.g. artificial neural networks (NNs) or hidden Markov models (HMMs)
- evolutionary information included (profiles)
- prediction accuracy $Q3 \approx 70\%$ (PhD, Rost 1993)

3rd generation methods

enhanced evolutionary sequence information (PSI-BLAST profiles) and larger sequence databases takes $Q3$ to $\approx 76\%$

PHD and PSIPRED are the best known methods

Current state-of-the-art in SSP

Mostly feed-forward neural-networks trained to predict H or E or C for each sequence position (residue) from windowed input:

```

EKGPDLYLIPLTEEAVAEAFYLAELRPRLRAEYALAPRKPAKGL EEALKRGAAGFAGFLGEDEL RAGEVTLKRLATGEQVRLSREEVPGYLLQALG
---VDIYLVASGADTQSAAMALAERLRDELkLMTNHGGGNFKKQF ARADKWGARVAVVLGESEVANGTAVVKDLRSGEQTAVAQDSVA AHLRTL LG
TKPKQMLVICLFEEALEELVWLAKLWREYNQVTIYPKVIKVDNGI RLANRLGYTFIGIVGKTD FDKKAITIKNLVSKQQT IYTWNELGERNV----
---VDVYMTAGEGTMMAGMKLAELRpGLRVMTHFGGGNFKKQF KRADKVGAAIALVLGEDEVAAQT VVVKDLAGGEQNTVAQAEVAKLL-----
-KGIDCYIVTLGEKAKDYSVSLVYKLR EaiSSEIDYENKKMKGQF KTADRLKARFIA ILGEDELAQNKINVKDAQ TGEQIEVALDEF-----
--TETQVFVATPQKNFLQERLKLIAELwsG IKAEMLYKNnkLLTQ LHYCESTGIPLVVI IGEQELKEGV I KIRSVASREEVA IKRENFVAEIQKRL
---TEVYVASAQKNLVRDRKKLVKMLRSaiKTEMALKAnkLLTQF QYAEERRIPLAIV IGEQELKDG VVKLRNVVTRDEQTIKLDQLITAVRDTL-
EEKEEVYFVIPFGDVHEYALRVADILRkKkVVEYSYRKGGLKKQL EFADKLGVKYAVI IGEDEVKNQEVTIKDMETGEQRRVKLSEL-----
---VEVYVASAHKGLHEQRLKVLNLLwaGVKAESHy1NPKLLVQL QHCEEHQIPLVVVLGDAELAQGLVKLREVT TREETNVKLEDLAAEIRR---
--TETQVFVATPQKNFLQERLKLIAELwsG IKAEMLYKNnkLLTQ LHYCESTGIPLVVI IGEQELKEGV I KIRSVASREEvrNRRDEV-----
---AKVLIACMHEEYFSYANRLAESLRQsiFSEVYPEAQKIKKPF SYANHKGHEFVAVIGEEEFKSETLSLKNMHSGMQLn1SFLKALEIIGE---
---PEVFVIPLKDMEKV-AINIAVKLreKI KTDIELSGRKL GKALDYANRVGAKLVIIVGKR DVERGVVTIRDMESGEQYNVSLNEIVDKVKNLL-

```

predicted:

CCCCEEEEEECHHHHHHHHHHHHHHHCCCCEEEECC~~C~~CCHHHHHHHHHHCCCCEEEECHHHHHCEEEEECCCCCEEECHHHHHHHHHHHHC

known:

CCCCEEEEECCHHHHHHHHHHHHHHHCCCCCEEECCCCCHHHHHHHHHHHCCCCEEEECHHHHHCEEEEECCCCCEEEECCHHHHHHHHHHHHC

$$Q_3 = \frac{\text{residues correct}}{\text{total residues}} \approx 76\%$$

(performance of predictors like PHD and PSIPRED)

Input encoding for SSP with NNs

Secondary structure “prediction” by homology

If sequence of unknown secondary structure has a homologue of known structure, it is *more accurate* to make an alignment and *copy the known secondary structure* over to the unknown sequence, than to do “ab initio” secondary structure prediction.

What is “known secondary structure”?

Of critical importance in training/assessment of SSP methods

Can be defined:

- visually by structural biologist
- by geometric and chemical criteria (ϕ , ψ angles, distances between atoms, hydrogen bonds...) by programs like DSSP and STRIDE

Other secondary structure prediction methods

- turn prediction
- transmembrane helix prediction
- coiled coil
- contact prediction, disulphides

What use is it?

No 3D means no clues to detailed function, so...

Accurate secondary structure predictions help sequence analysis: finding homologues, aligning homologues, identifying domain boundaries.

Can help true 3D prediction

Further improvements to SSP...?

Long range information

Folding pathway and/or 3D information...