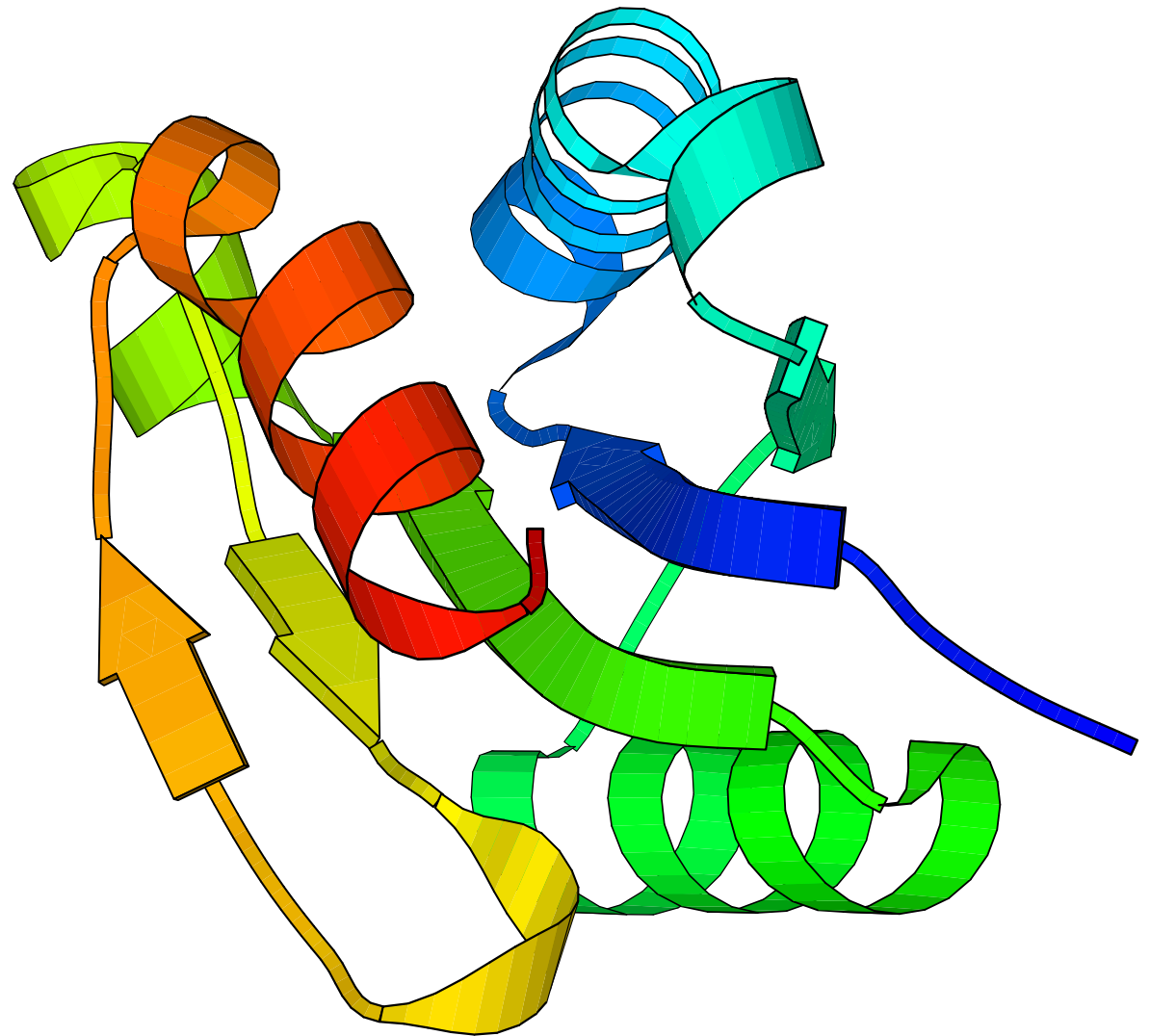


Protein Structure Prediction

Bob MacCallum, Stockholm Bioinformatics Center

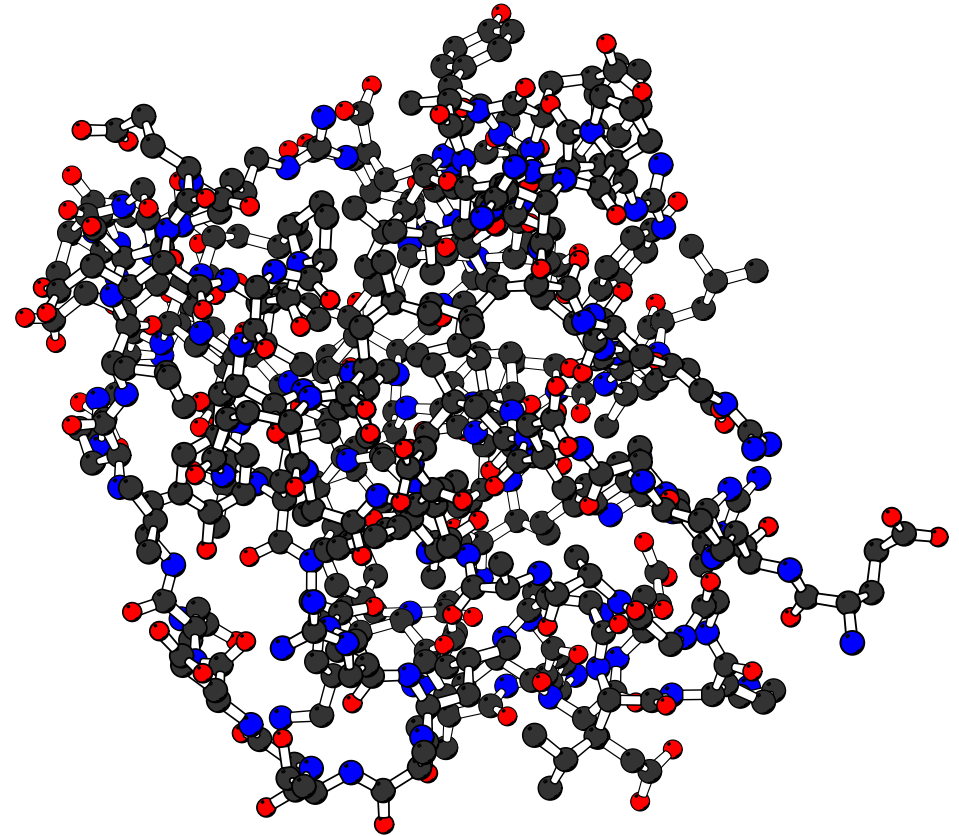
```
EKGPDLYLIPLTEEAVAEAF  
YLAEALRPRLRAEYALAPRK  
PAKGLEEALKRGAAGFLG  
EDEL RAGEVTLKRLATGEQV  
RLSREEVPGYLLQALG
```

+ 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

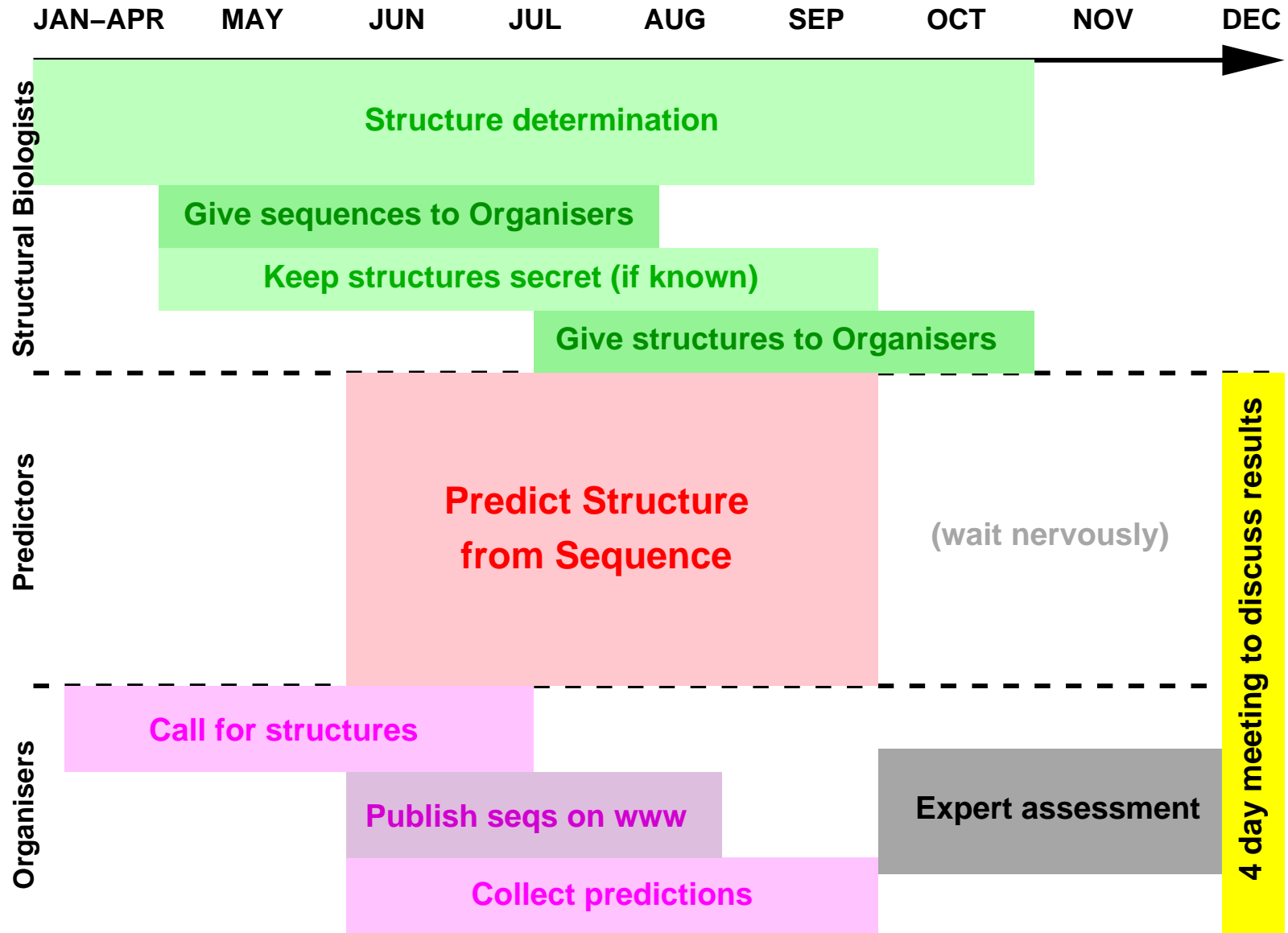
Critical Assessment of Techniques for Protein Structure Prediction

Why do we have CASP?

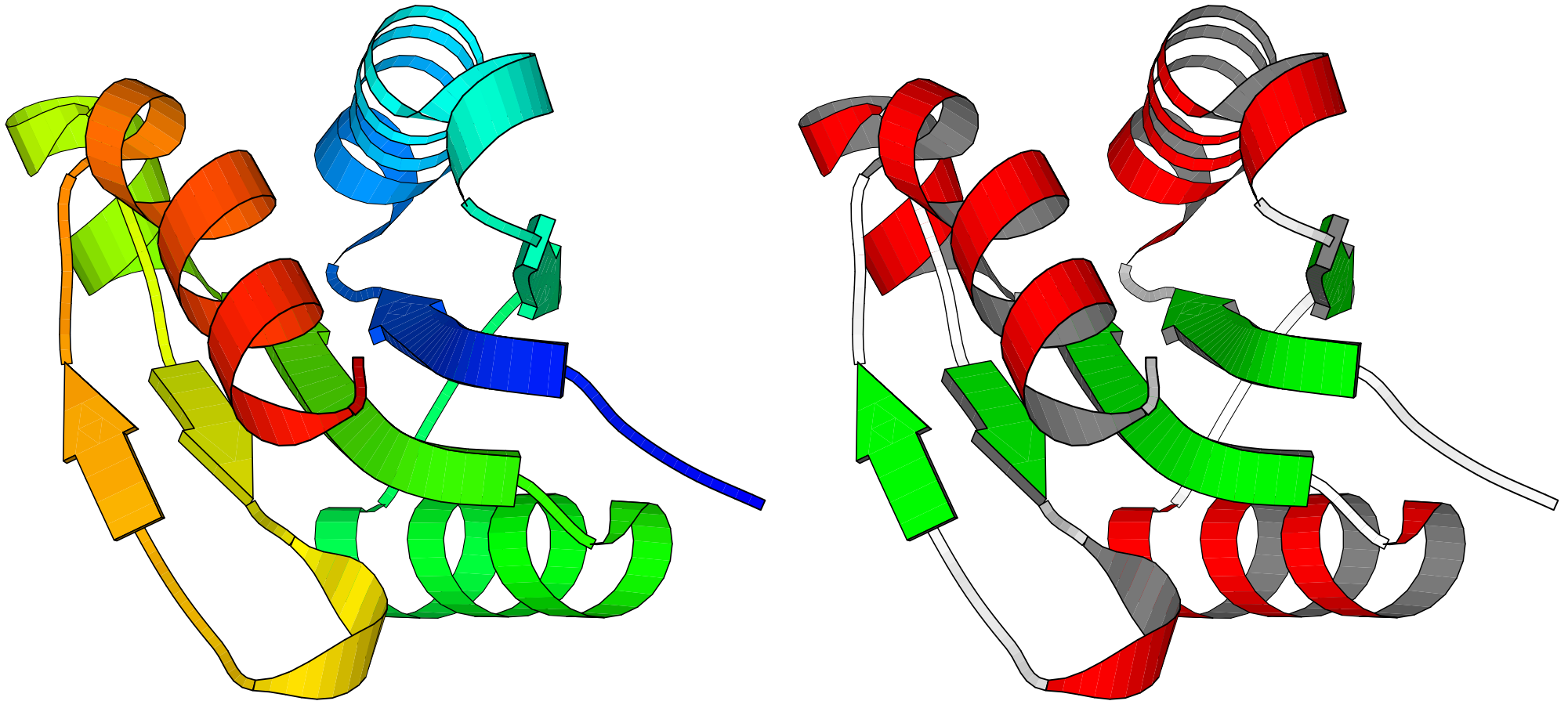
- People cheat!
 - people work hard to make prediction programs work for their favourite proteins, but...
 - benchmarking may be polluted by “information leakage”
- Difficult to compare methods fairly
 - software and data issues
 - different measures, standards

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 RLANRLGYTFIGIVGKTDFDKKAITIKNLVSKQQTIIYTWNELGERNV----
---VDVYMTAGEGTMMAGMKLAEQLrpGLRVMTHFGGGNFKKQF KRADKVGAAIALVLGEDEVAAQTVVVKDLAGGEQNTVAQAEVAKLL-----
-KGIDCYIVTLGEKAKDYSVSLVYKLR EaiSSEIDYENKKMKGQF KTADRLKARFIAILGEDELAQNKINVKDAQTGEQIEVALDEF-----
--TETQVFVATPQKNFLQERLKLIAELwsG IKAEMLYKNnkLLTQ LHYCESTGIPLVVIIGEQLKEGVIKIRSVASREEVAIKRENFVAEIQKRL
---TEVYVASAQKNLVRDRKKLVKMLRSaiKTEMALKAnkLLTQF QYAEERRIPLAIVIGEQLKDG VVKLRNVVTRDEQTIKLDQLITAVRDTL-
EEKEEVYFVIPFGDVHEYALRVADILRkKkVVEYSYRKGGLKKQL EFADKLGVKYAVIIGEDEVKNQEVTIKDMETGEQRRVKLSEL-----
---VEVYVASAHKGLHEQRLKVLNLLwaGVKAHSy1NPKLLVQLQHCEEHQIPLVVVLGDAELAQGLVKLREVT TREETNVKLEDLAAEIRR---
--TETQVFVATPQKNFLQERLKLIAELwsG IKAEMLYKNnkLLTQ LHYCESTGIPLVVIIGEQLKEGVIKIRSVASREEvrNRRDEV-----
---AKVLIACMHEEYFSYANRLAESLRQsiFSEVYPEAQKIKKPF SYANHKGHEFVAVIGEEEFKSETLSLKNMHSGMQLn1SFLKALEIIGE---
---PEVFVIPLKDMEKV-AINIAVKLreKI KTDIELSGRKL GKALDYANRVGAKLVIIVGKRDVERGVVTIRDMESGEQYNVSLNEIVDKVKNLL-

```

predicted:

CCCCEEEEEECHHHHHHHHHHHHHHHCCCCEEEECC~~C~~CCHHHHHHHHHHCCCCEEEECHHHHHCEEEEECCCCCEEEECHHHHHHHHHHHHC

known:

CCCCEEEEEECHHHHHHHHHHHHHHHCCCCCEEECCCCCHHHHHHHHHHHCCCCEEEECHHHHHCEEEEECCCCCEEEECHHHHHHHHHHHHC

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

(performance of predictors like PHD and PSIPRED)

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

Homology-based modelling

“Comparative modelling”

Homology = common ancestry or evolutionary relatedness.

Pairs of homologous protein domains almost always have similar 3D structure. Differences mainly in loops connecting secondary structures.

- find and **align** homologue(s) of known structure (parent)
- copy **core** backbone coordinates of parent(s) to unknown
- try to add **loops** from database or *ab initio*
- add **sidechains** and minimise errors/clashes

Finding parent(s) for homology modelling

Usually done with BLAST, PSI-BLAST or similar.

Models are typically 1-3Å rmsd using parents > 40% sequence identity.
Lower sequence identity → poorer quality, less useful models.

Alignment from BLAST may be used in modelling program

Multiple parents can be used

Building the core model

Following the alignment, copy the backbone co-ordinates of the template into the model

Insertions and deletions are hopefully in the loops connecting core secondary structures

Trim back loop take-off points

(This is the easy bit)

Loop building

Problem: the loops in your target sequence are different lengths and/or sequence to the loops in the parent. Loops are known to be the most variable part of protein structures.

Two approaches:

library based look for loops in known structures with same number of residues and same take-off point geometry

ab initio build brand-new loop from one take-off point to the other; this may involve fragments of known structure from a library.

Mutually contacting loops are a problem. Ideally you should build them all simultaneously, but this is slow.

Tidying up the model

Some or all of the model may still be backbone atoms only. So sidechains are added in the most probable conformations, and to minimise clashes.

Energy minimisation can be used to “relax” the model to give more reasonable backbone and side-chain geometry. Basically a short run of *molecular dynamics*.

Often the best models come from modelling programs which make the *fewest* changes to the template structure. The quality is largely dependent on the *alignment quality* and *choice of parent(s)*. Hence energy minimisation can make a model worse!

Manual inspection (perhaps correcting the alignment or changing the parents) and repeating the process can help too.

Homology Modelling on the Web

SwissModel runs as a web service

Modeller is a program you download

WHAT-IF is a more like a graphical interface for protein structure manipulation, and you can make homology models too

For more details on homology modelling (from a WHAT-IF perspective):
<http://www.cmbi.kun.nl/gvteach/hommod/>

Homology Modelling at CASP

Some improvements in alignment quality and building large loops

No progress at the atomic-level (refinement often gives a worse model)

Fold recognition (FR)

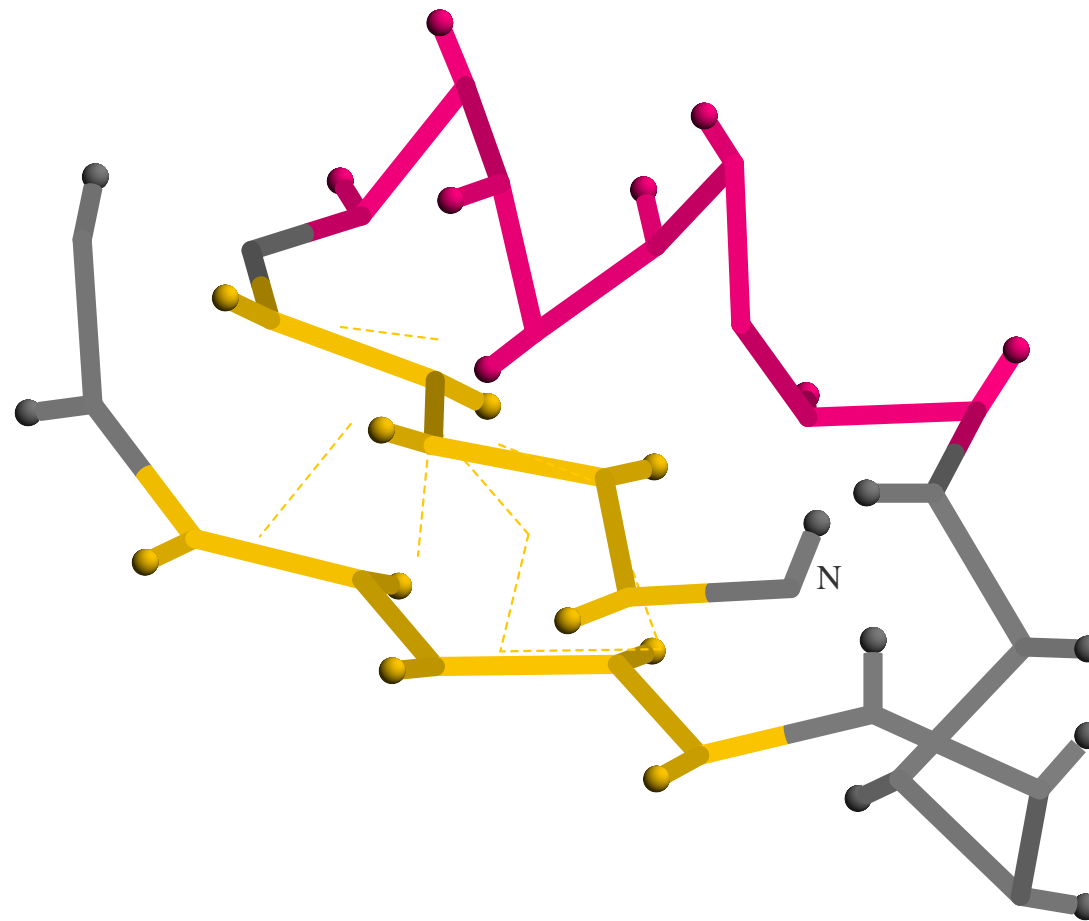
- aligning sequences to known structures (the “fold library”)
- like homology modelling except you’re lucky if you can find a parent!
- used when ‘standard’ sequence methods fail
- generally uses structural information
- *very* rough models can be made

Basic principles of FR

- structure conserved more than sequence
- structural constraints on sequence
 - locally – i.e. sec. str. preferences, Gly/Pro in turns
 - globally – hydrophobic core, residue contacts
- sequence-structure alignment must make sense in 3D
 - no gaps in core secondary structures
 - no missing strands from sheets

FR methodologies

Information	Method name(s)	Approach
Amino-acid sequence only	PDB-BLAST, SAM-T99, SUPERFAMILY, ...	Profile-based or hidden Markov model-based alignments, often against a specially prepared library
Sequence and predicted secondary structure	3D-PSSM, INBGU, FUGUE, ...	Structure-enhanced sequence alignment (matching predicted to real secondary structure, structural environments)
Sequence and 3D-structure	GenThreader, THREADER, ...	Alignment quality is evaluated in terms of pairwise residue contacts
Other FR servers	Pcons, ShotGun, ...	Consensus of different FR results. Basically when different methods agree, you can be more sure...



VIFVLWGNAARQKCNLLFQTKHQHAVLACPH

3D profiles

The original sequence-structure approach. Doesn't really work, but historically interesting...

[your notes here]

Bowie JU, Lüthy R, Eisenberg D. *A method to identify protein sequences that fold into a known three-dimensional structure.* Science. 1991 Jul 12;253(5016):164-70

Pair potentials

Outline: thread the sequence onto a known backbone structure, optimise the alignment so that the *intramolecular* side-chain interactions are most favourable.

With the double-dynamic programming alignment algorithm of THREADER and the Bryant method (refs. below) it's possible to ignore sequence information from the known structure.

Other so-called 'frozen approximation' pair-potential methods use it.

The THREADER paper (short on details, unfortunately):

Jones DT, Taylor WR, Thornton JM. *A new approach to protein fold recognition*. Nature. 1992 Jul 2;358(6381):86-9.

Another successful method using a different alignment algorithm:

Panchenko A, Marchler-Bauer A, Bryant SH. *Threading with explicit models for evolutionary conservation of structure and sequence*. Proteins. 1999;Suppl 3:133-40.
<http://www.ncbi.nlm.nih.gov/Structure/RESEARCH/casp3/pap.html>

GenThreader

Original THREADER takes way too long. GenThreader takes short-cuts:

- fast sequence-sequence alignment against fold library
- build 3D-model from alignment
- evaluate pair potentials in model
- evaluate solvent potentials in model
- train a neural network to make decision based on: alignment score, sequence lengths, pair-potential score, solvation score

Jones DT. *GenThreader: an efficient and reliable protein fold recognition method for genomic sequences*. J Mol Biol. 1999 Apr 9;287(4):797-815

Jones DT, Tress M, Bryson K, Hadley C. *Successful recognition of protein folds using threading methods biased by sequence similarity and predicted secondary structure*. Proteins. 1999;37(S3):104-111.

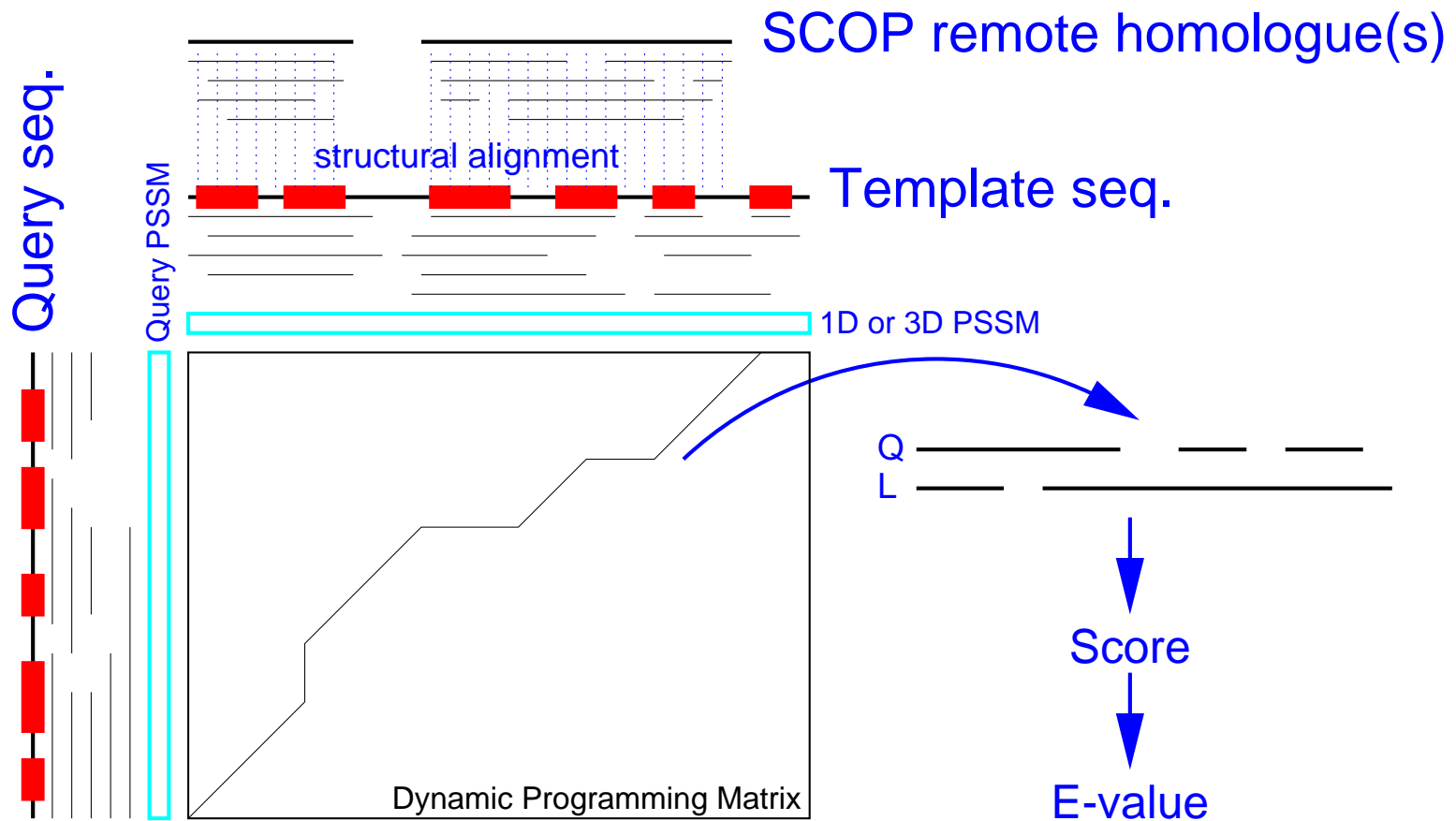
Profiles and secondary structure matching

ca. 1994 secondary structure prediction accuracy was respectable

simple FR methods matching sequence and secondary structure did quite well

in CASP2&3, careful use of PSI-BLAST was competitive with FR methods

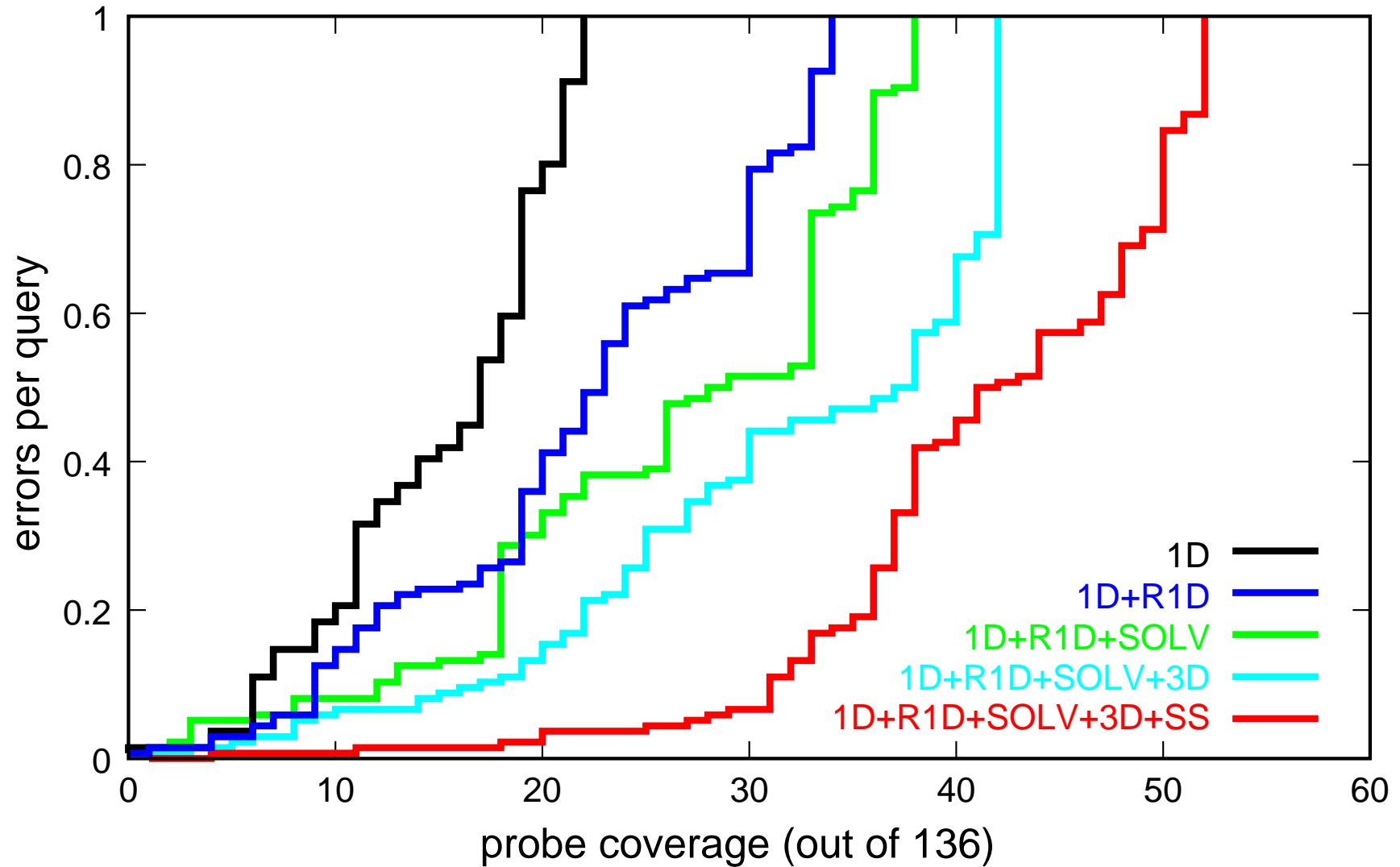
3D-PSSM methodology



Design criteria: should run fast, use multiple seqs., predicted sec. str and other structural information

Kelley LA, MacCallum RM, Sternberg MJ. *Enhanced genome annotation using structural profiles in the program 3D-PSSM*. J Mol Biol. 2000 Jun 2;299(2):499-520.

3D-PSSM benchmarking



Probe-template pairs: same SCOP superfamily, PSI-BLAST $E > 100$

Practical aspects of FR

- Pre-processing of sequences
 - ignore non-globular regions: transmembrane regions, coiled coil, low complexity regions, signal peptides
 - identify domain boundaries & repeats - run separately
 - make best possible multiple sequence align./sec. str. pred.
- consensus between methods
- can you trust confidence values?
- compare function of query (if known) with templates (SAWTED)
- check models make sense in 3D
- is function conserved between query and template
- hand edit alignments?

How good is FR?

LiveBench and CASP measure performance.

E-values work reasonably well.

In the real world, you might get a few percent more “hits” with FR compared to PSI-BLAST.

Individual researcher vs. genome-wide analysis

Structure information not necessary?

Tertiary structure prediction

- “true” *ab initio*
- three main approaches
 - simulation: fold up polypeptide from extended form
 - screening: make many structures, select “best”
 - fragment assembly: “ROSETTA” hybrid local prediction, simulation & screening

None of these really work “off the shelf”, although ROSETTA has been applied to all Pfam families of unknown structure.

Rosetta outline

Best new fold (ab initio) method in CASP3&4

- Start with extended chain
- Monte Carlo fragment assembly
- Repeat MC many times (and for homol seqs)
- Filter models
- Cluster models
- Pick large clusters

Monte Carlo optimisation

1. Initial configuration (random or extended)
2. Make a randomised MOVE on configuration
3. Measure change in quality of structure (DE)
 - (a) IF better ($DE < 0$) ACCEPT MOVE
 - (b) ELSIF $\text{rand}(1) < e^{-DE/kT}$ ACCEPT MOVE
 - (c) ELSE REJECT MOVE
 - (d) GO TO 2. (reduce T if you like)

Rosetta MC Energy Function

Compactness (radius of gyration)

Hydrophobic burial

Polar side chain contacts (statistical pairwise potential)

Hydrogen bonding between beta-strands

Hard-sphere repulsion (VdW)

Rosetta: Filtering the models

Between 6,000 and 150,000 models generated

Contact Order

- Generated models are biased towards simple structures
- Filter models to give correct contact order distribution for domains of that size/composition

Sheet filter

Add side chains, calculate atomic physical potential (to eliminate poorly packed structures)

Rosetta: clustering the models

Compare models to each other with RMSD

Models can come from different family members

Cutoff varied to give 80-100 members in largest cluster

The largest clusters are assumed to contain the best structures
(attractors in folding space...?)

Summary of lectures

Structure prediction methods are continuously being developed/improved

Most successful methods use known structures (modelling, fold recognition)

New folds methods are beginning to work for smaller proteins

Structural genomics may get most new folds

Intellectual challenge remains

Protein design, nanotechnology...

Globular proteins are only part of the story