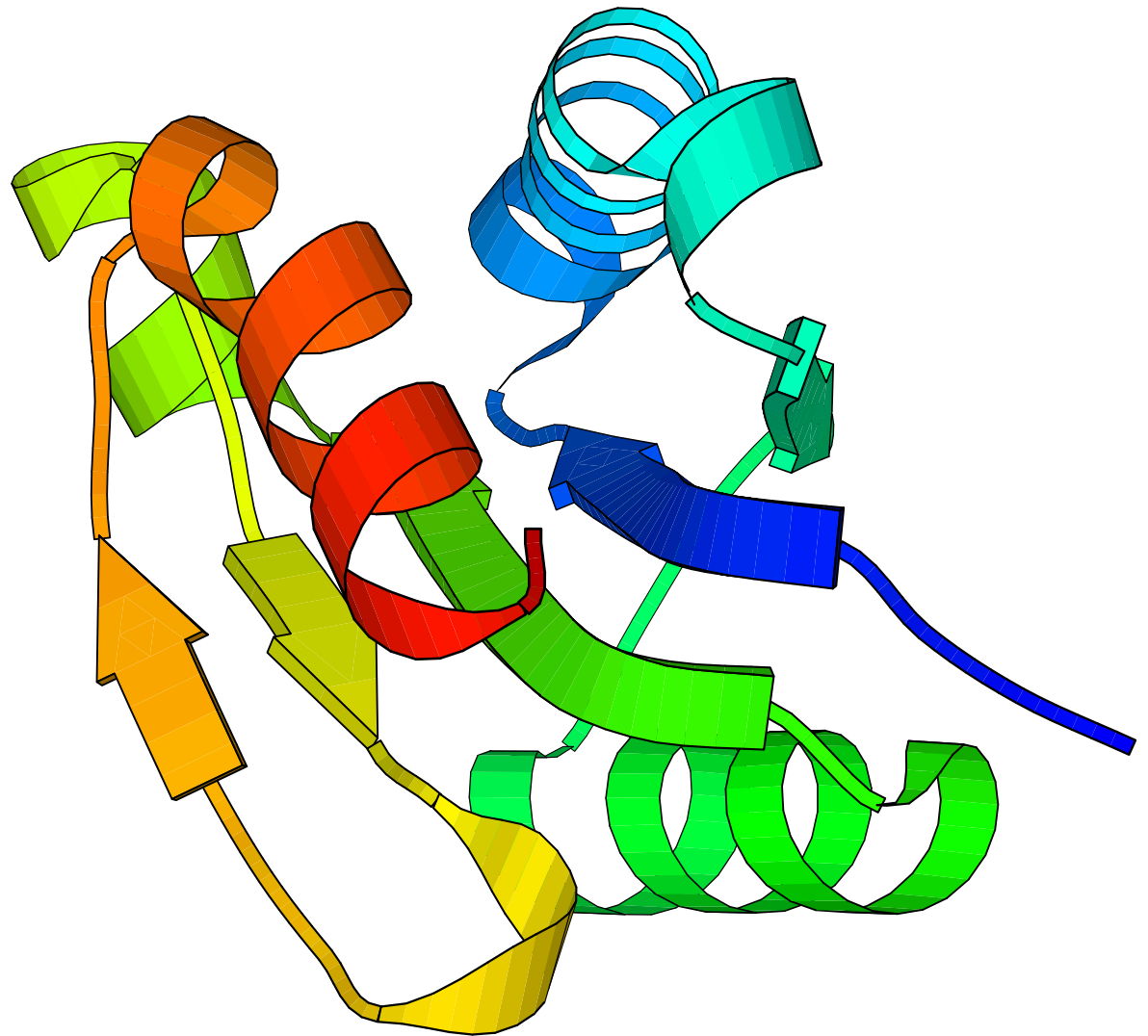


Protein Structure Prediction II

Bob MacCallum, Stockholm Bioinformatics Center

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

+ computer =



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

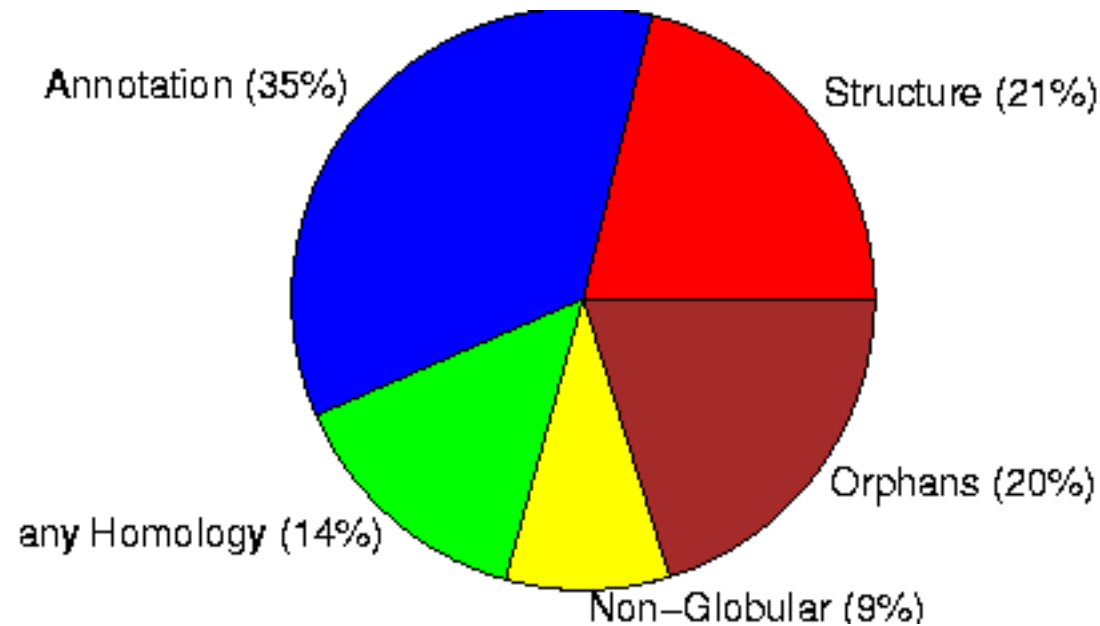
Structural genomics

- *ab initio* methods not very good
- homology based methods ok
- limited structure universe
- \$\$\$ + robots = structures of current “white space”
- skip difficult proteins/structures

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

Drosophila genome



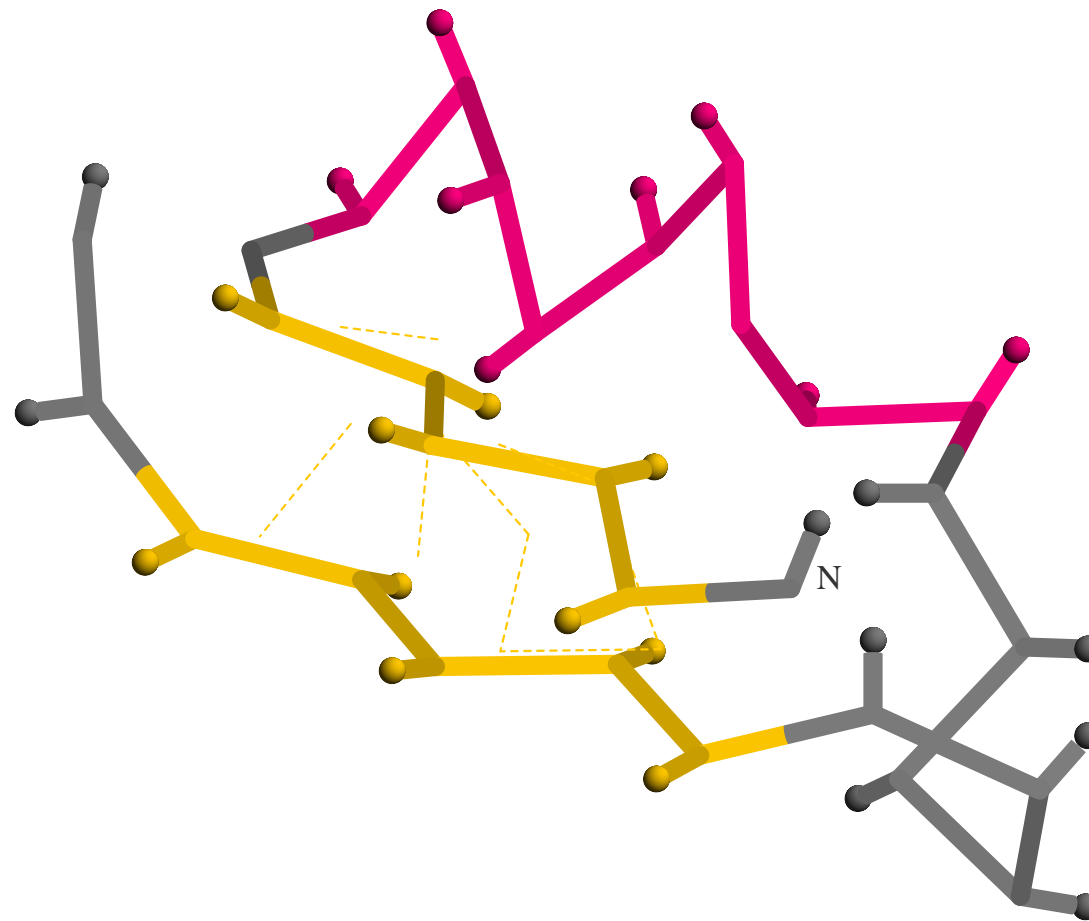
- 21% assigned to structure (PSI-BLAST)
- should be possible to assign an estimated 20% more to remote homologues of known structure

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.

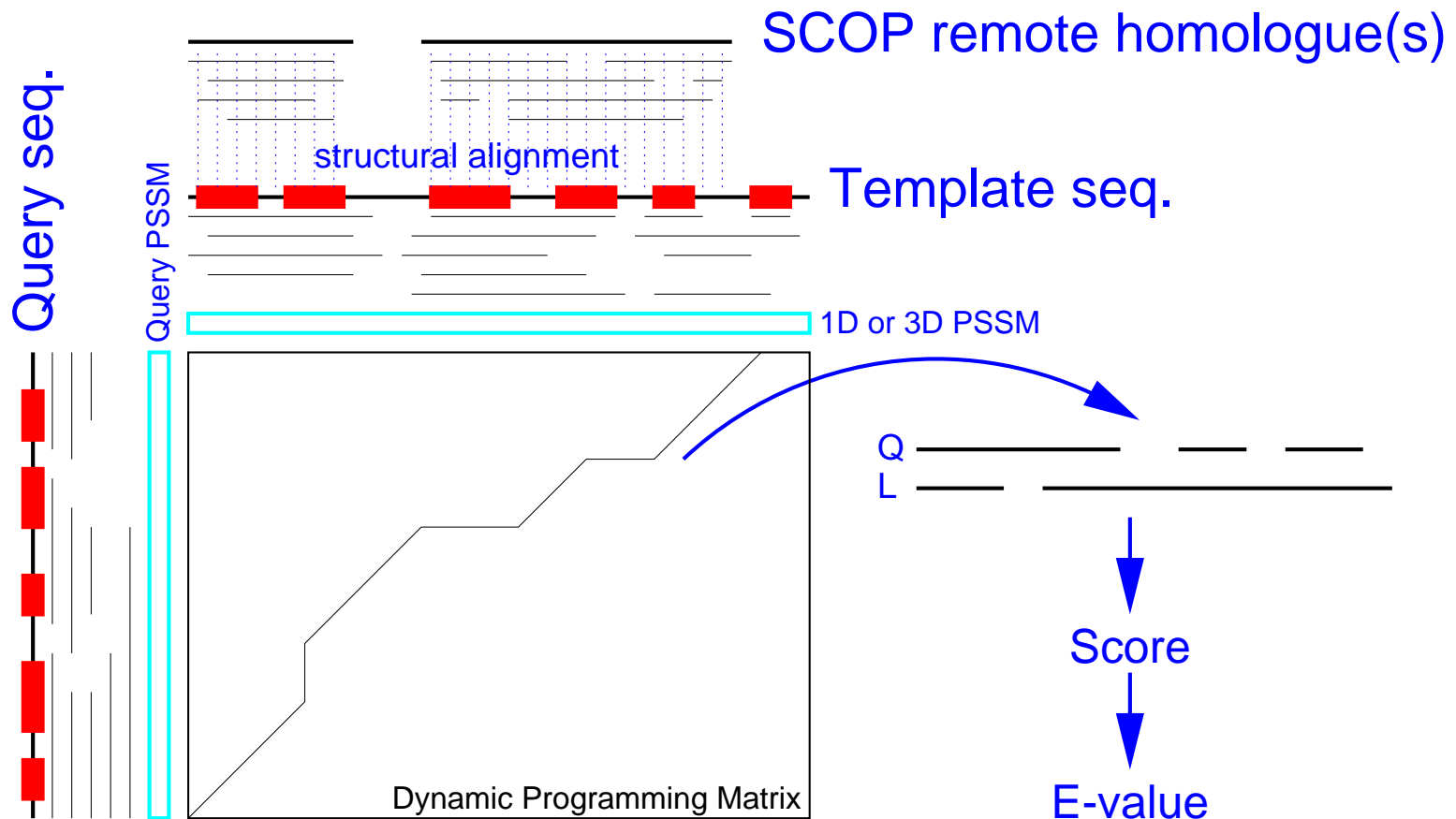
Profiles and secondary structure matching

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

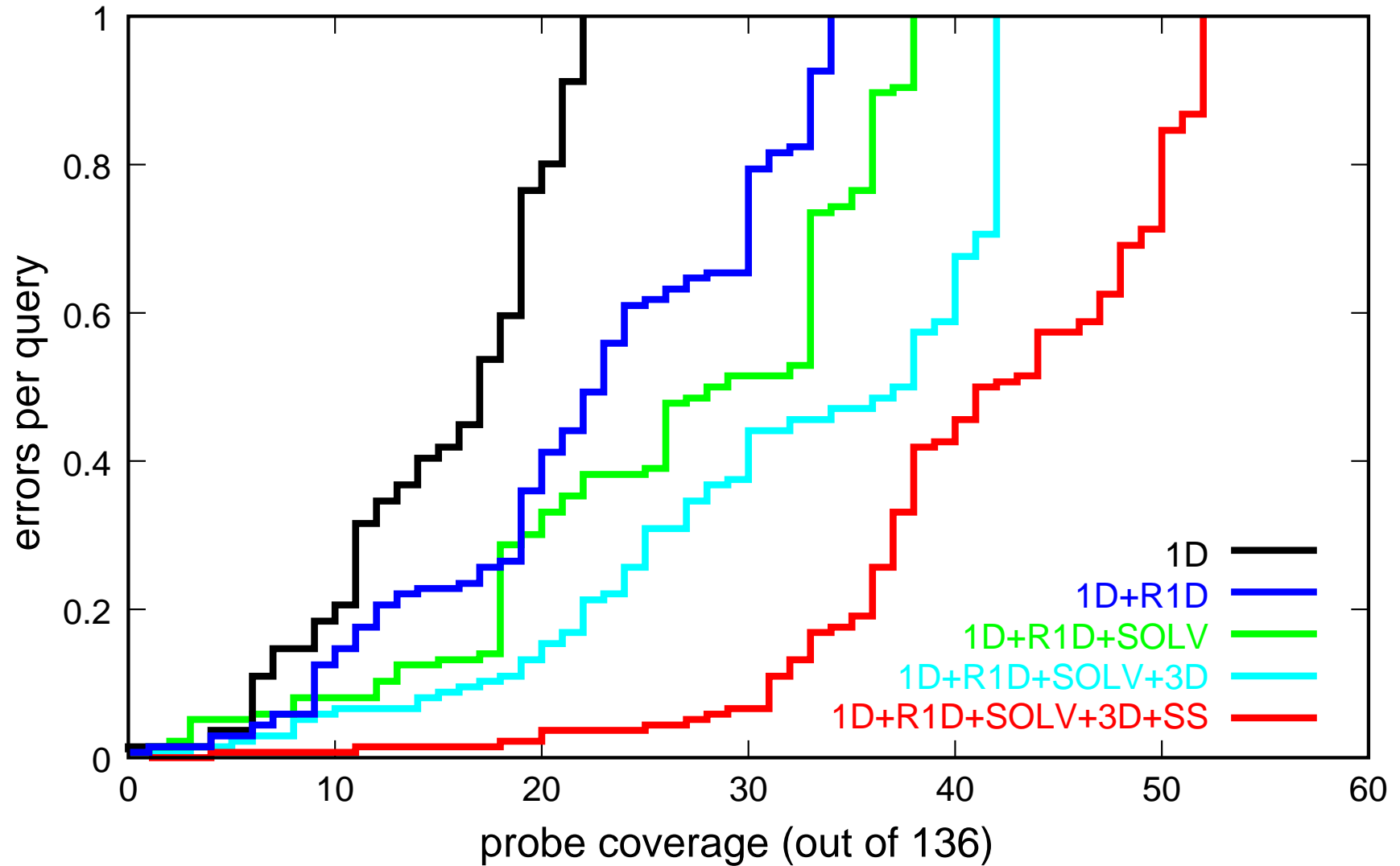
3D-PSSM methodology



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

Fragment assembly

Don't try to predict every detail of the backbone/sidechains

Use fragments of known structures

In Rosetta they are 3 and 9 residues long

Selected by a PSI-BLAST search against PDB sequences with a generous E-value threshold

Other people have used fragment assembly too

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, and tested at CASP/LiveBench/EVA

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

New folds methods are beginning to work for smaller proteins

But fold recognition is OK if structural genomics gets all folds

We haven't talked about protein design

We've only talked about globular proteins