7.91 Amy Keating

Methods for Protein Structure Prediction

Homology Modeling &

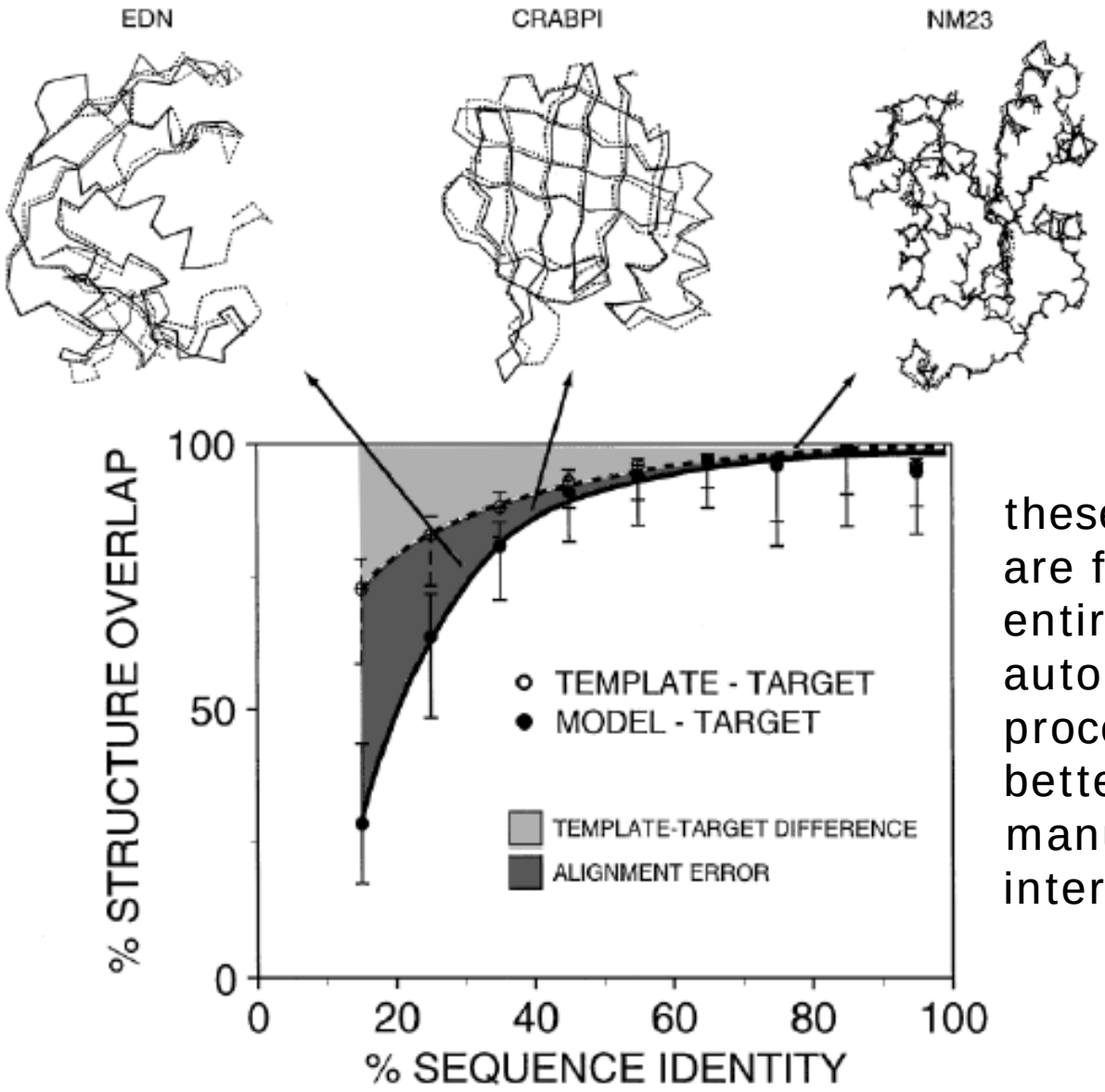
Fold Recognition

Next time: Ab Initio Prediction

Review - **Homology Modeling**

- Identify a protein with similar sequence for which a structure has been solved (the *template*)
- Align the target sequence with the template
- Use the alignment to build an approximate structure for the target
- Fill in any missing pieces
- Fine-tune the structure
- Evaluate success

An excellent review:
Marti-Renom et al. *Annu. Rev. Biophys. Biomol. Struct.* 29 (2000): 291-325.



these numbers are from an entirely automated process - can do better with manual intervention

Marti-Renom et al. *Annu. Rev. Biophys. Biomol. Struct.* 29 (2000): 291-325.

Homology Modeling on a Genomic Scale

- Requires automation
  - Can't choose templates or fine-tune the alignment by hand!

- MODBASE and 3D-CRUNCH

  http://alto.compbio.ucsf.edu/modbase-cgi/index.cgi

  http://www.expasy.ch/swissmod/SM_3DCrunch.html

- Automatic assessment is critical - how reliable is the model?

One approach to assessment

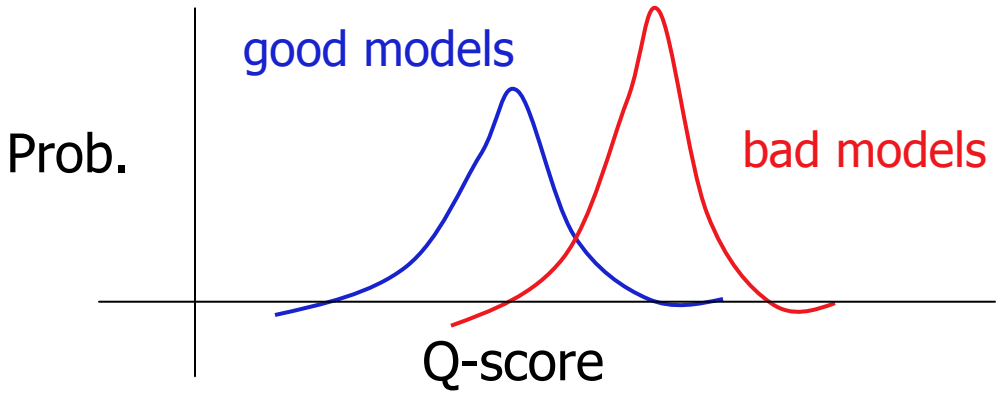
Want to compute the probability that a prediction is good, based on properties of the model

For a given score of the model (e.g. Q-score - more on this later), use a training set of known examples, together with Bayes' rule

$$P(A|B) = P(A \wedge B)/P(B) = P(A)P(B|A)/\{P(A)P(B|A) + P(!A)P(B|!A)\}$$

Assume probability of a good vs. a bad model is the same,
i.e. $P(A) = P(!A)$ where A = good model; !A = bad model; B = Q-score

$$P(good|Q\text{-}score) = P(Q\text{-}score|good)/\{P(Q\text{-}score|good) + P(Q\text{-}score|bad)\}$$



Sanchez, R, and A Sali. "Large-scale Protein Structure Modeling of The Saccharomyces Cerevisiae Genome." *Proc Natl Acad Sci U S A.* 95, no. 23 (10 November 1998): 13597-602.

MODBASE

http://alto.compbio.ucsf.edu/modbase-cgi/index.cgi

- 733,239 sequences & 7,120 non-redundant structures
- **Fold Assignments (by PSI-BLAST)**
- Reliable fold assignments: 827,007 for 413,311 sequences
- Average folds per sequence: 2.0
- Average length of queries: 511 amino acids
- Average length of folds: 229 amino acids
- **Comparative Models (by MODELLER)**
- Reliable models 547,473
- Sequences with reliable models: 327,393 (59%)
- Structures used as templates: 6.366 (89%)

For a reliable fold assignment, PSI-BLAST E value < 0.0001
OR a reliable model.
For a reliable model, 30% of Cα atoms superpose within 3.5Å of their correct positions

Example

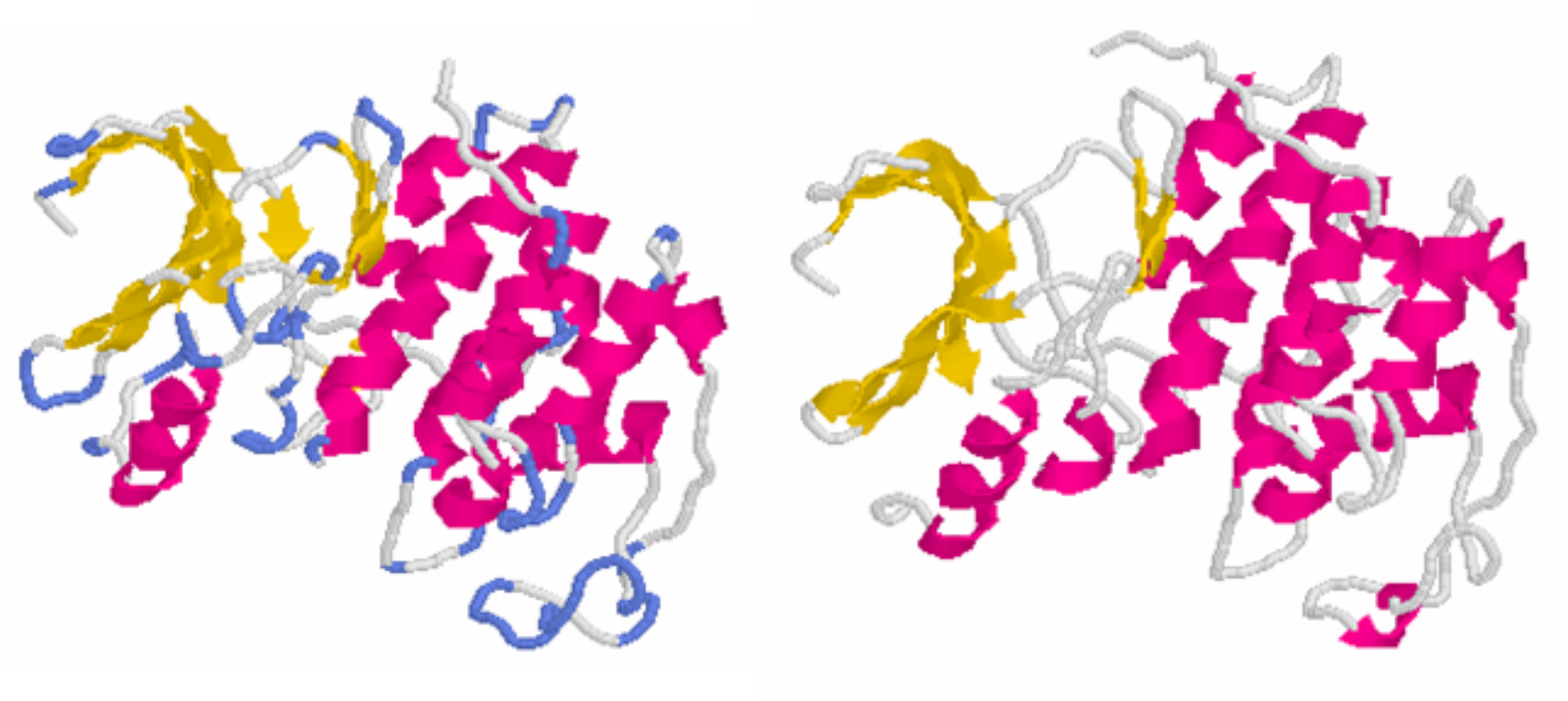
You've just cloned a new gene from Pombe - look it up in ModBase

- putative galactosyltransferase associated protein kinase (GenBank accession # 3006192)

TARGET						MODEL DATA					TEMPLATE			
Model/Fold Reliability	Sequence based View	Sequence Database Links	Database Annotation	Organism	Protein Size	Modeled Segment	Size	Seq Id (%)	E-value	Model Score	PDB code	Template based View	Segment	Annotation
●		TR Q60145	serine/threonine protein kinase Dataset: SP/TR-2001 PFAM PRODOM	*Schizosaccharomyces pombe*	398	71-368	298	45.00	1e-121	1.00	1hcl		1-291	human cyclin-dependent kinase 2
●		TR Q60145	serine/threonine protein kinase Dataset: SP/TR-2001 PFAM PRODOM	*Schizosaccharomyces pombe*	398	55-396	342	30.00	1e-111	1.00	1p38		2-342	map kinase p38
●		TR Q60145	serine/threonine protein kinase Dataset: SP/TR-2001 PFAM PRODOM	*Schizosaccharomyces pombe*	398	45-397	353	23.00	3e-71	1.00	1kob A		2-324	twitchin
●		TR Q60145	serine/threonine protein kinase Dataset: SP/TR-2001 PFAM PRODOM	*Schizosaccharomyces pombe*	398	32-397	366	22.00	3e-63	1.00	1apm E		2-328	$c-/amp$-dependent protein kinase (e.c.2.7.1.37) (c/apk) 1apm 3 2 (catalytic subunit) "alpha" isoe

Pieper, Ursula, Narayanan Eswar, Ashley C. Stuart, Valentin A. Ilyin, and Andrej Sali. "MODBASE, A Database of Annotated Comparative Protein Structure Models." *Nucl. Acids Res.* 30 (2002): 255-259.
http://alto.compbio.ucsf.edu/modbase-cgi/index.cgi

Model of new POMBE gene



TARGET

TEMPLATE = 1HCL

PDB ID: 1HCL

Schulze-Gahmen, U., J. Brandsen, H. D. Jones, D. O. Morgan, L. Meijer, J. Vesely, and S. H. Kim. "Multiple Modes of Ligand Recognition: Crystal Structures of Cyclin-dependent Protein Kinase 2 in Complex with ATP and Two Inhibitors, Olomoucine and Isopentenyladenine." *Proteins* 22 (1995): 378.

The CASP contests

- **C**ritical **A**ssessment of Protein **S**tructure **P**rediction
- Began in 1994 (CASP1)
- Held every two years
- Experimentalists submit target sequences
- Predictors submit and rank blind predictions
- Assessors develop criteria to judge success
- A meeting is held to discuss the results and a journal issue (of PROTEINS) is published to describe them
- In theory, this identifies the problem areas and people go back and work on them for the next round of CASP

CASP4 Target T0111

Example of a CASP target

1. Protein Name
   enolase
2. Organism Name
   Escherichia coli
3. Number of amino acids (approx)
   431
4. Accession number
   P08324
5. Sequence Database
   Swiss-prot
6. Amino acid sequence
   SKIVKIIGREIIDSRGNPTVEAEVHLEGGFVGMAAAPSGASTGSREALEL
   RDGDKSRFLGKGVTKAVAAVNGPIAQALIGKDAKDQAGIDKIMIDLDGTE
   NKSKFGANAILAVSLANAKAAAAAKGMPLYEHIAELNGTPGKYSMPVPMM
   NIINGGEHADNNVDIQEFMIQPVGAKTVKEAIRMGSEVFHHLAKVLKAKG
   MNTAVGDEGGYAPNLGSNAEALAVIAEAVKAAGYELGKDITLAMDCAASE
   FYKDGKYVLAGEGNKAFTSEEFTHFLEELTKQYPIVSIEDGLDESDWDGF
   AYQTKVLGDKIQLVGDDLFVTNTKILKEGIEKGIANSILIKFNQIGSLTE
   TLAAIKMAKDAGYTAVISHRSGETEDATIADLAVGTAAGQIKTGSMSRSD
   RVAKYNQLIRIEEALGEKAPYNGRKEIKGQA
7. Additional Information
   oligomerization state: dimer in the presence of magnesium by dynamic light scattering
   and small angle x-ray solution scattering and
   in the recently solved crystal structure.
8. **Homologous Sequence of known structure**
   yes
9. Current state of the experimental work
   Structure solved by molecular replacement. Currently,
   the refinement to 2.5 A resolution is near completion.
   Current Rfree 27 % ; R 22 %

BLAST target T0111 against the PDB

```
>gi|1311141|pdb|1PDZ|   Mol_id: 1; Molecule: Enolase; Chain: Null; Synonym:
          2-Phospho-D-Glycerate Dehydratase; Ec: 4.2.1.11;
          Heterogen: Phosphoglycolate; Heterogen: Mn 2+
 gi|1311142|pdb|1PDY|   Mol_id: 1; Molecule: Enolase; Chain: Null; Synonym:
          2-Phospho-D-Glycerate Dehydratase; Ec: 4.2.1.11
         Length = 434

 Score =  384 bits (987), Expect = e-107
 Identities = 220/432 (50%), Positives = 280/432 (63%), Gaps = 16/432 (3%)

Query: 3   IVKIIGREIIDSRGNPTVEAEVHLEGGFVGMAAAPSGASTGSREALELRDGDKSRFLGKG 62
           I K+  R I DSRGNPTVE +++   G    AA PSGASTG  EALE+RDGDKS++ GK
Sbjct: 3   ITKVFARTIFDSRGNPTVEVDLYTSKGLF-RAAVPSGASTGVHEALEMRDGDKSKYHGKS 61

Query: 63  VTKAVAAVNGPIAQALI--GKDAKDQAGIDKIMIDLDGTENKSKFGANAILAVSLANAKA 120
           V  AV  VN  I   +I  G     Q   D+ M  LDGTENKS  GANAIL VSLA  KA
Sbjct: 62  VFNAVKNVNDVIVPEIIKSGLKVTQQKECDEFMCKLDGTENKSSLGANAILGVSLAICKA 121

Query: 121 AAAAKGMPLYEHIAELNGTPGKYSMPVPMMNIINGGEHADNNVDIQEFMIQPVGAKTVKE 180
            AA  G+PLY HIA L     +  +PVP  N+INGG HA N + +QEFMI P GA +  E
Sbjct: 122 GAAELGIPLYRHIANL-ANYDEVILPVPAFNVINGGSHAGNKLAMQEFMILPTGATSFTE 180

Query: 181 AIRMGSEVFHHLAKVLKAK-GMN-TAVGDEGGYAPNLGSNAEALAVIAEAVKAAGYELGK 238
           A+RMG+EV+HHL  V+KA+ G++ TAVGDEGG+APN+ +N +AL +I EA+K AGY  GK
Sbjct: 181 AMRMGTEVYHHLKAVIKARFGLDATAVGDEGGFAPNILNNKDALDLIQEAIKKAGYT-GK 239
```

etc…

Best prediction for T0111 at CASP4 superimposed with the real structure

For a description of results from CASP 4 homology modeling, see…

Tramontano, A, R Leplae, and V Morea. "Analysis and Assessment of Comparative Modeling Predictions in CASP4." *Proteins* Suppl 5 (2001): 22-38.

Progress in Comparative Modeling

Methods have not advanced significantly from CASP1 to CASP5

More template structures are available

More sequences are available to help alignment

More remotely related sequences can be detected using PSI-BLAST

No new good solutions to the alignment OR refinement problem

The fold recognition / threading approach to protein structure prediction

OBSERVATION: there appear to be a limited number of protein folds (~1,000?)

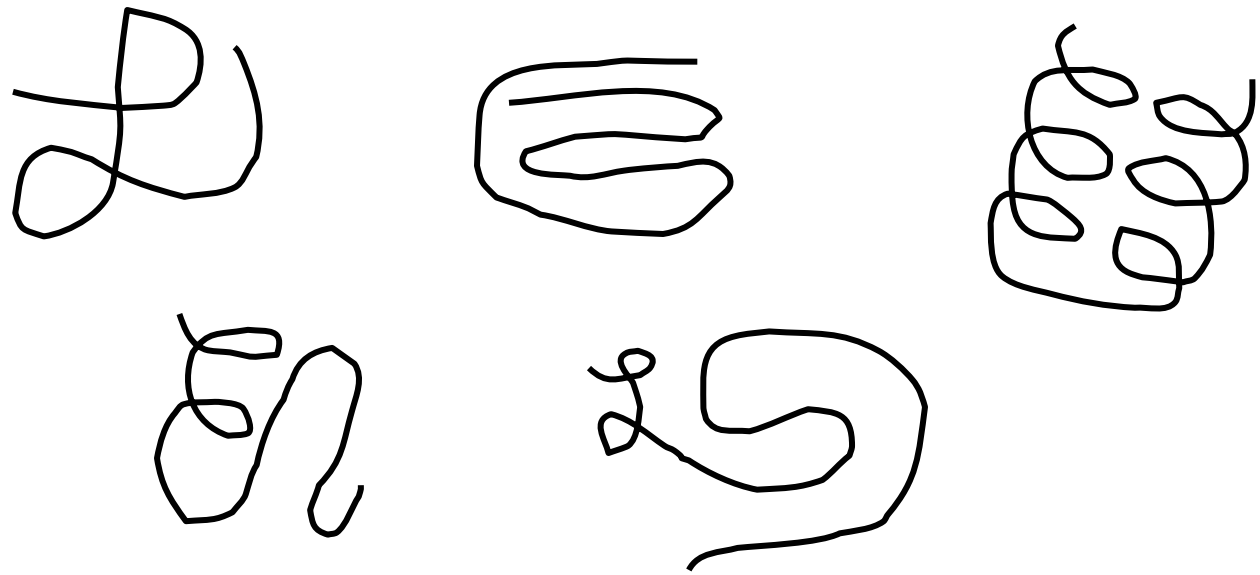
Instead of having to predict protein structure “from scratch”, maybe we can just pick the correct answer out of a finite list

This can be done using sequence-based techniques, or by “threading” the sequence onto different templates in turn, and evaluating how good a match each one is

Fold recognition or threading

Target = SHPALTQLRALRYCKEIPALDPQLLDWLLLEDSMTKRFEQQ...

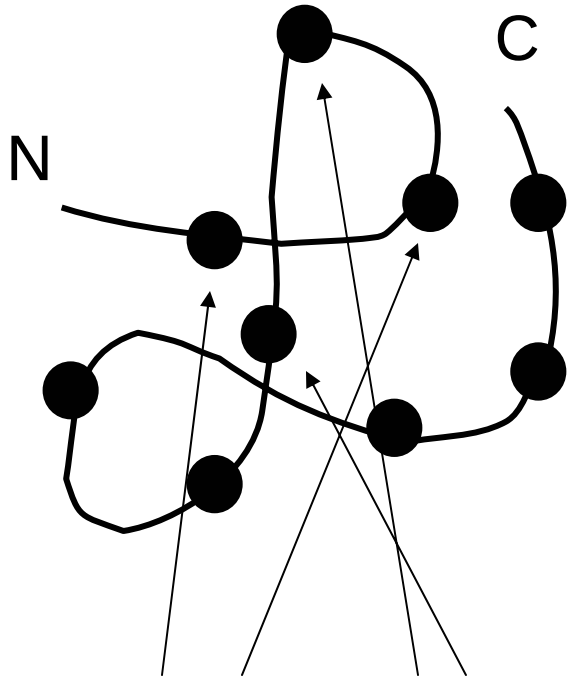
Library of possible folds
(these have known sequences AND structures):



Sequence-structure alignment

Target = SHPALTQLRALRYCKEIPALDPQLLDWLLLEDSMTKRFEQQ...

= $t_1 t_2 t_3 t_4 t_5 \ldots t_n$



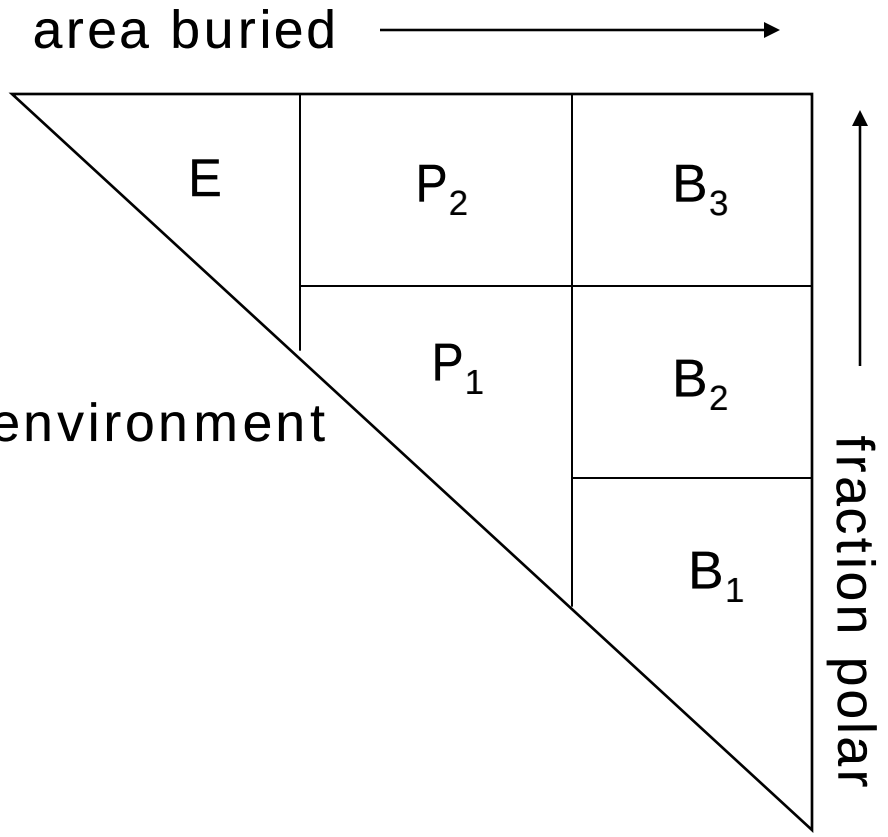
Sequence for known fold = $s_1 s_2 s_3 s_4 s_5 \ldots s_n$

Positions for known fold = $p_1 p_2 p_3 p_4 p_5 \ldots p_n$

How do you align the target sequence to the structure?

Linking the sequence to structural properties by 3D-1D comparison

- Describe the structure by a sequence of terms representing the structural environment of each residue



–How buried it is

–Polar/non-polar nature of the environment

–Local secondary structure

6 x 3 environment classes

Bowie & Eisenberg, Science (1991) 253, 164-170

Different amino acids prefer different environments

- Quantify preference of each amino acid type for each environment using statistical preferences (log odds score)

$$score_{ij} = \ln\left(\frac{P(j\_in\_environment\_i}{P(j\_in\_any\_environment}\right)$$

environment class	Trp	Phe	Tyr	...
B1α	1.00	1.32	0.18	...
B1β	1.17	0.85	0.07	...
⋮				

Make a scoring matrix = 3D profile

fold position	environ. class	Trp	Phe	Tyr	...	gap
1	B1b	1.17	0.85	0.07	...	200
2	E loop	-2.14	-1.90	-0.94	...	2

and use it to align the sequence to the environment string using dynamic programming

target sequence	$p_1\ p_2\ p_3\ p_4\ p_5\ p_6\ p_7$	*environment class*
t_1	. . .	
t_2		
t_3		
t_4		
...		

Fold recognition by 3D-1D

- Compare the target sequence alignment to the template against a large number of other possible sequences

$$Z_{score} = \frac{score - < score >}{\sigma}$$

- Z-scores > 7 represent a good match

Improvements to 3D-1D scoring

- Better to use more classes - this is possible now that we have a lot more structural data
- Incorporate predicted properties of the target (i.e. 2° structure)
- H3P2 uses 5 scoring dimensions
  - 3 for the fold
    - 7 residue classes
    - 3 secondary structures
    - 2 burial groups
  - 2 for the sequence
    - 7 residues classes
    - Predicted secondary structure
- 7x3x2x7x3 = 882 different elements in the scoring matrix
- Derive values for the matrix from 119 structurally similar pairs with < 30% sequence identity

H3P2 method: Rice & Eisenberg J. Mol. Bio. (1997) 267, 1026

Fold recognition by 3D-1D alignment

Advantages

Disadvantages

Fold recognition by 3D-1D alignment

Advantages
- fast $O(mn)$
- incorporates structural information
- reasonable performance

Disadvantage
- assumes independence of positions
- assumes conservation of environment

Useful both for fold recognition and for structure assessment (e.g. of predicted or experimental structures)

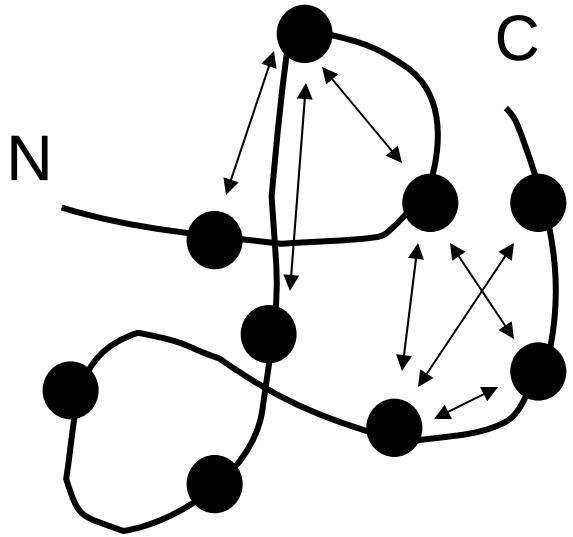
Incorporating position-dependence

- Score based on a pair-wise contact potential

$$Score = \sum_{i} \sum_{j>i} score(i, j)$$

$$score(i, j) = f(p_i, p_j, t_{r_i}, t_{r_j})$$

t_{r_i} is the amino acid from the target sequence that is mapped to structure position i



Knowledge-based contact potentials

- Use observed frequencies in the pdb to compute scores

Example

Define a contact as occurring if 2 residues are < 6 Å apart (Cα-Cα distance)

$$score(i,j) = -\ln\left(\frac{P(i,j \mid contact)}{normalization}\right)$$

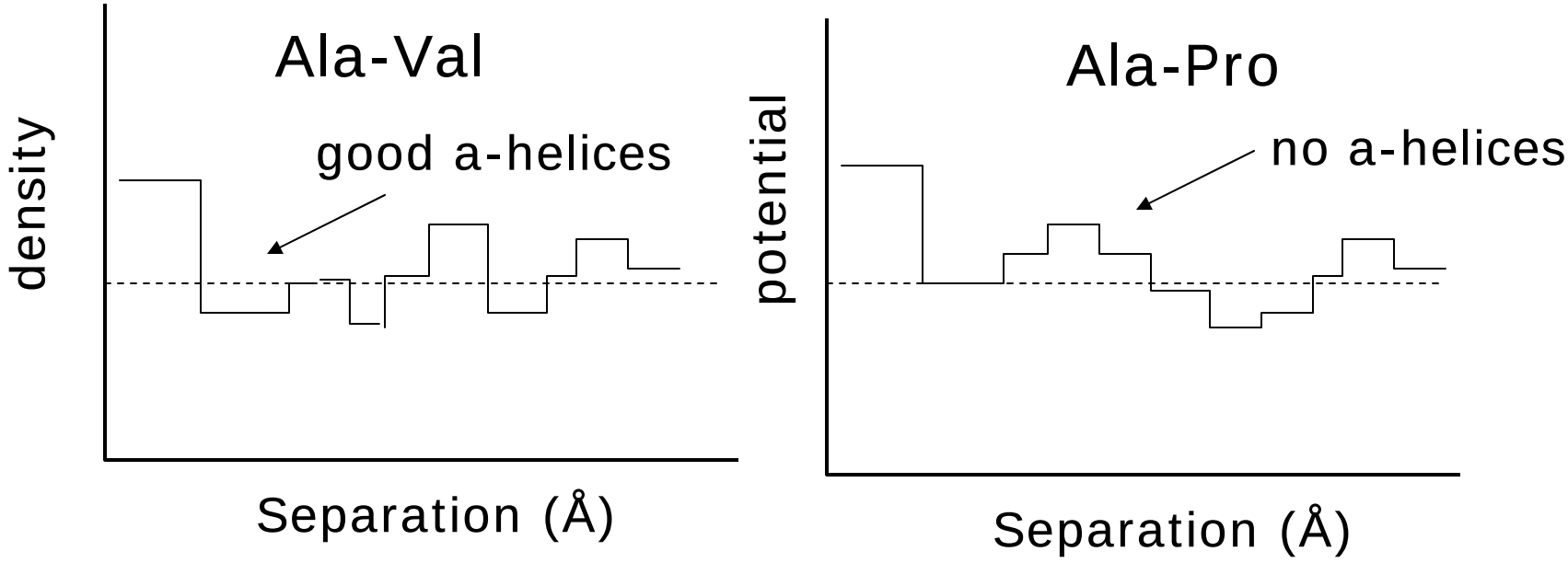
Normalization based on the expected rate of seeing i and j in contact, given no interaction between the two.

Knowledge-based threading potentials

- Some statistical potentials include a distance-dependence

$$score(aa_i, aa_j, r_{ij}, d_{ij}) = -\ln\left(\frac{f(aa_i, aa_j, r_{ij}, d_{ij})}{f(r_{ij}, d_{ij})}\right)$$

At d_{ij} = 4 compare potentials for



Sippl J. Mol. Bio. (1990) 213, 859; Jones et al. Nature (1992) 358, 86

Pros and cons of contact potentials

Pros and cons of contact potentials

- Fast to compute
- Not sensitive to details of structure
- Can use even for low-resolution experimental structures
- Don’t require accurate description of physics
- Have proven to be quite sensitive to quality of structure

- Don’t represent physical potentials well
- Tend to capture mostly H/P patterning effects
- Artifacts: +/+, +/- and -/- are similarly good at distances > 4Å since they are often all found on the surface

Using contact potentials for threading or structure evaluation

Sippl defined a “polyprotein” of 230 proteins of known structure fused together with reasonable geometry

Slide the target sequence along the polyprotein and compute a Z-score; normalize somehow for the length

$$Z_{score} = \frac{score - < score >}{\sigma}$$

This is the Q-score used by ModBase to compute model reliability. It is independent of the scoring functions used to build the models.

Problem with using contact potentials for threading

- The contacts depend on the alignment
- The alignment depends on the contacts

To calculate the score for putting a residue in a certain position, you need to know what residues are in other positions. These aren't yet determined!

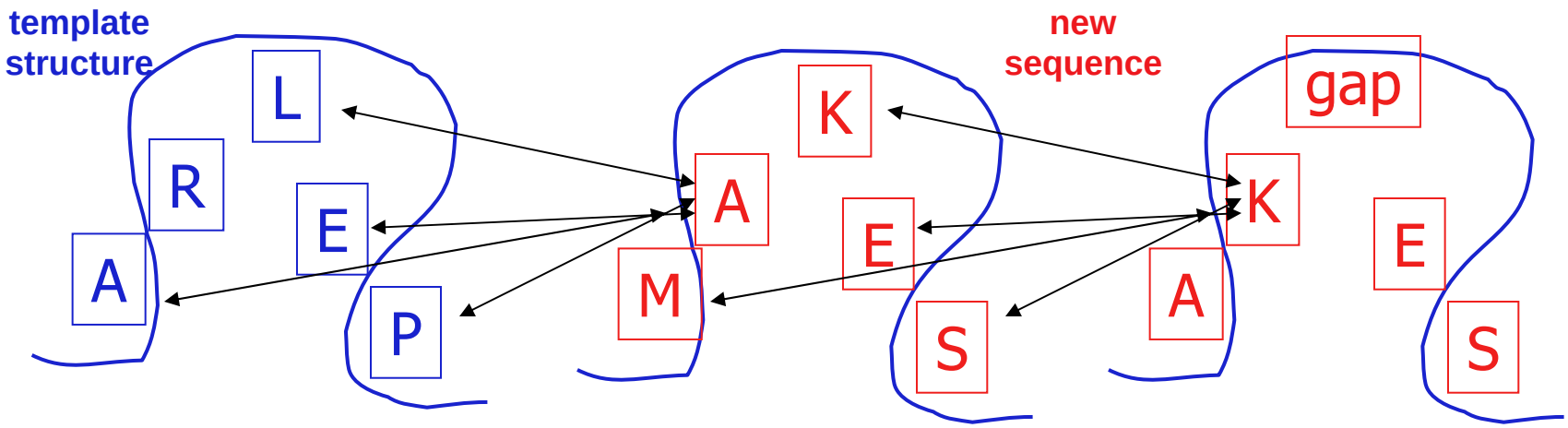
Performing an alignment using a pairwise scoring function while allowing variable-length gaps is an NP-hard problem - it can't be solved in polynomial time

What to do?

- Put limits on gap lengths and positions (e.g. don't allow gaps in core secondary structure elements)
- Use heuristics

  Example: in the "frozen" approximation you first use the template sequence to compute the scores at each position

  In subsequent iterative rounds you use the residue that was there in the last round of alignment



Godzik et al. J. Mol. Bio. (1992) 227, 227-238

Fold recognition performance - CASP4

- Two tasks
  - Find the correct fold
  - Align the target to the template
- Difficulty is correlated with **how similar the best template is to the target** and **how similar the target sequence is to a template sequence**
- For the best groups, they usually recognize the correct fold (or something close)
- For the worst groups performance is terrible (worse than the performance of automated servers)
- For all groups, alignment is A HUGE PROBLEM!!

Sippl et al. PROTEINS (2001) Suppl. 5, 55-67

Fold recognition performance CASP4

VERY POOR

GOOD
(but only 9% residues
correctly aligned)

EXCELLENT
(46% residues
correctly aligned)

Please see

Sippl, MJ, P Lackner, FS Domingues, A Prlic, R Malik, A Andreeva, and M Wiederstein. "Assessment of The CASP4 Fold Recognition Category." *Proteins* Suppl 5 (2001): 55-67.

Fold recognition at CASP4

Scale:

1 = found somewhat related fold

2 = found right fold

3 = right fold, poor alignment

4 = SUPER! (still, alignment accuracy ~40%)

Average performance over targets:

Homolog (sim. struct. & funct. in pdb)	analog (sim. struct in pdb)	new fold (part of struct in pdb)
3.7	2.6	1.7
2.5	0.8	0.9

First line: “virtual predictor” averages best score from any group
Second line: average for best group

Sippl et al., PROTEINS (2001) 5, 55-67

BEST TEMPLATE
12.7% seq ID

TARGET

BEST PREDICTION

Please see

Kinch, LN, JO Wrabl, SS Krishna, I Majumdar, RI Sadreyev, Y Qi, J Pei, H Cheng, and NV Grishin. "CASP5 Assessment of Fold Recognition Target Predictions." *Proteins* 53, Suppl 6 (2003): 395-409.

Assessment criteria at CASP

Complicated area of research - makes it hard to follow progress in the field as the criteria keep changing

Recent consensus that GDT-TS is a good measure

GDT-TS = 1/4(N1 + N2 + N3 + N4)

N1 = max # residues alignable to w/in	1 Å rms
N2 =	2 Å
N3 =	4 Å
N4 =	8 Å

Target Difficulty

Please see

Venclovas, C, A Zemla, K Fidelis, and J Moult. "Assessment of Progress over The CASP Experiments." *Proteins* 53, Suppl 6 (2003): 585-95.

Fold recognition at CASP5

Fold recognition performance in CASP5 improved primarily because of the use of “metaservers”

Metaservers collect predictions from other methods and combine them in different ways (e.g. using neural networks)

Some metaservers:
3D SHOTGUN
PCONS

Fold recognition on a genome-wide scale

- Want to annotate various proteomes for structure and function
- The threading methods are too slow and require too much human intervention for genome-wide applications
- Sequence-based methods have gotten very good
- Adding structural information helps in detecting remote homologies

Programs for genome-wide fold recognition

- **GenThreader** http://bioinf.cs.ucl.ac.uk/psipred
  - Build a structure-based sequence alignment from all the fold templates
  - Align the target to the profile (*sequence* alignment, like PSI-BLAST)
  - Score the alignment using a threading potential

$$E(aa_i, aa_j, d_i) = -\ln\left(\frac{f(aa_i, aa_j, r_{ij}, d_{ij})}{f(r_{ij}, d_{ij})}\right) \qquad E_{environ}(a_i) = -\ln\left(\frac{f^{ai}(burial)}{f(burial)}\right)$$

  - Get out several measures of success:
    - Alignment score, alignment length, target length, template length, pairwise threading score, environment threading score
  - Feed these to a neural network to get a single indicator of the quality of the model

Performance of GenThreader

- Benchmark on 68 protein pairs with < 18.9% sequence identity from FSSP (remember DALI…)
- 73.5% of matches made correctly
  - Best sequence-based methods in 1999 got 63%
- Low false positive rate - good indication of confidence
- 46.2% of residues correctly aligned when fold was correct
- Mycoplasma genitalium genome (1999)
  - Provided some annotation for 46% of proteins in the genome

    (30% of amino acids)

Jones, David T. "GenTHREADER: An Efficient and Reliable Protein Fold Recognition Method for Genomic Sequences1." *Journal of Molecular Biology* 287, no. 4 (9 April 1999): 797-815.