Protein Fold Recognition
remote homologues
sequence to structure alignments

Andrew Torda, wintersemester 2008 / 2009, AST

Aims

- sequence with no close homologues of known structure
  - find related proteins
  - try to build a model

  - not necessarily a good model

Remote homologues

- sequence similarity is not enough

Main problems

- not easy to identify remote homologues
- alignments will not be reliable

Consequence

- less emphasis on sequence-based methods
  - sequence to structure methods

- scheme …

What to do

Given a sequence

- can one find homologues of known structure ? (simple blast)
  - alignment to homologues
  - build models

  stop
- look for remote homologues - psi-blast, sensitive methods
  - careful alignment to homologues
  - build models

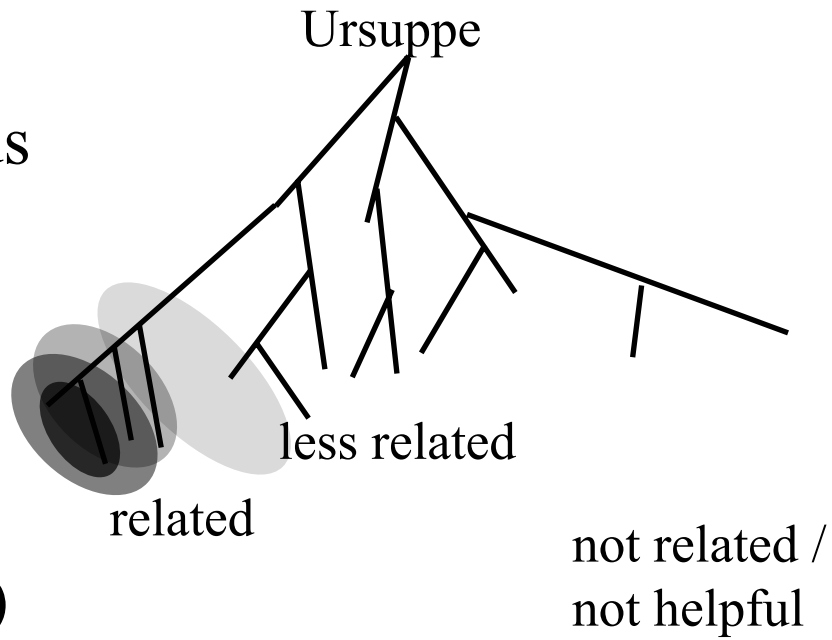
  stop
- desperation
  - what we talk about now
- first - summary

alignment methods

	slow	fast
methods	Needleman & Wunsch / Smith-Waterman	seeded – blast, fasta, suffix tree methods
time	O(nm) or O(nm^2) (sequence sizes)	O(nk) – database size
guaranteed to find optimal solution	yes	no
very remote homologues	may work	less likely to work

- How far can one go with sequence based methods ?
  - try iterative methods

Remote homologues from sequence



- Still prefer sequence based methods

psi-blast philosophy

- for my protein
  profile is original sequence
  while (known structure not found)
    find more close homologues
    use for average sequence "profile"

- how many loops can you do ?
  - each brings in more proteins – some wrong proteins

Alternative philosophy

- Database methods are fast / approximate

- Would we do better with more careful alignments
  - use only sequences from the protein data bank
    - 6×10^4 instead of millions
  - careful full dynamic programming style alignments
    - Smith and Waterman / Needleman & Wunsch
  - substitution matrix for remote homologues

- would give better alignments
- might help find some templates (better scoring alignments)

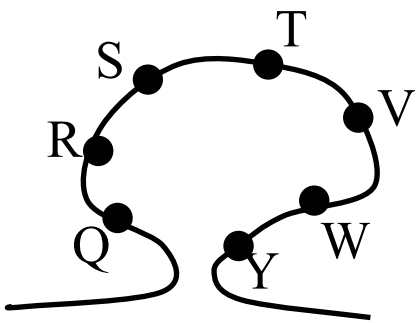
Slower alignments – not really used

- careful alignments might help – not done in practice
  - change philosophy
- From sequence viewpoint

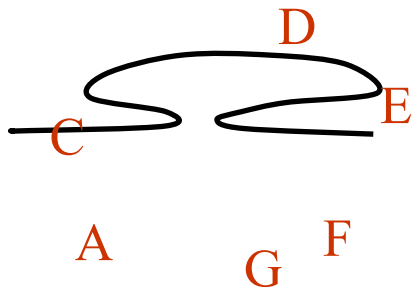
```
..AC-DEFG..
..QRSTVWY..
```

- When aligning to known structures, more information

```
..AC-DEFG..   query sequence
..QRSTVWY..   known structure
```



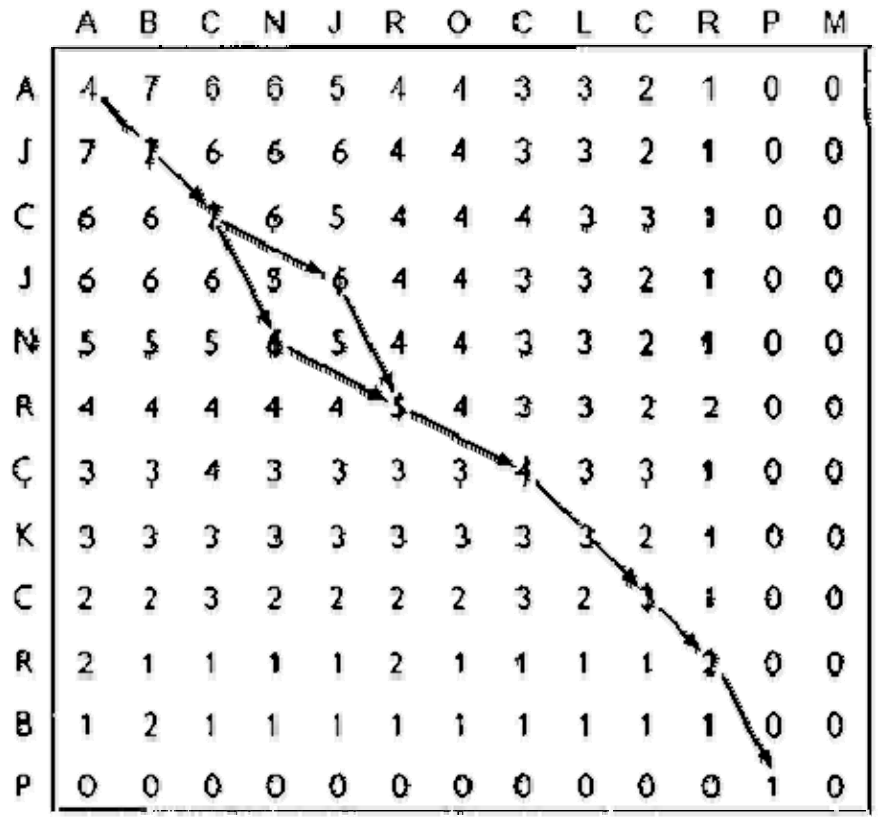
known structure



model implied

Sequence to structure alignments

- Remember how sequence alignments work
  - similarity / substitution scores
  - fill out score matrix
  - find best path
- Can we use this for sequenc to structure alignments ?



Needleman, SB & Wunsch, CD, (1970) J. Mol. Biol. 48, 443-453

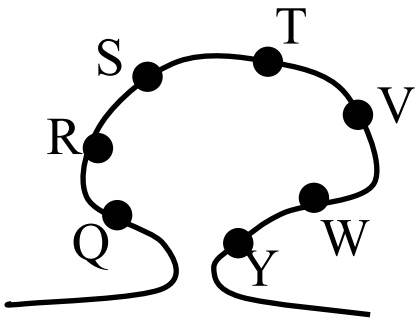
more exotic scoring

- From sequence viewpoint

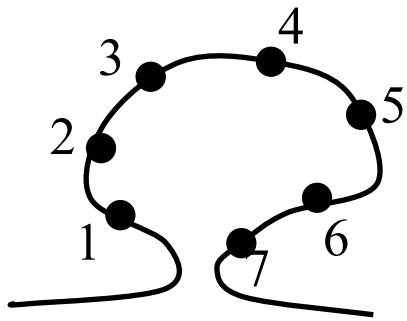
  `..AC-DEFG..` my sequence

  `..QRSTVWY..` a protein of known structure

- rather than just align sequences, could I use the structure ?



known structure



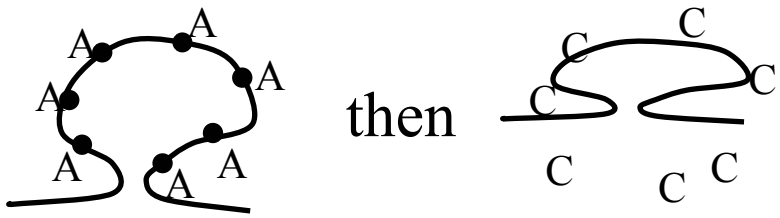
forget sequence

- score matrix ?

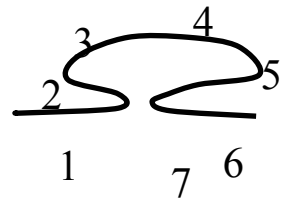
	A	C	D	E	F	G
1	?	...				
2						
3						
4						
5						
6						
7						

sequence to structure scoring

- I have to be able to place (A, C, D..) at each position and get a suitability score



	A	C	D	E	F	G
1	?	...				
2						
3						
4						
5						
6						
7						



- then it would be easy to do sequence to structure alignments
- advantage:
  - we claim that structure is more conserved than sequence
  - can find appropriate/fitting/suitable structures for a sequence
  - very remote, but reliable homologues
- vorsicht !!!!

sequence to structure scoring

- define an energy function
  - depends on interaction of residue with structure
    - easy
  - depends on interaction with neighbours
    - but who are the neighbours ?

A, C, D, or E …

- bad news
  - we cannot even fill out a column in the score matrix
  - to test every combination of neighbours
    - NP-complete
- an excuse to try some approximations

	A	C	D	E	F	G
1		?				
2		?				
3		?				
4		?				
5		?				
6		?				
7		?				

approximations for scoring

- two problems
  - we do not know where all the atoms are – side chain coordinates
  - to score "C" at each position we need to know neighbours
- side-chains : ignore / average
  - use a score / energy function which averages over all conformations
- neighbour positions : much harder
  - environment description
  - frozen approximation

Environment description

- an example of profiles (case study)
- we know
  - certain sites are hidden from solvent (middle of protein)
    - only compatible with trp, phe, ile, … (hydrophobic)
  - some sites are involved in "salt bridges"
  - some secondary structures are preferred by certain residues
- can one count the probabilities of residue types ?

- overview
  - collect list (parameterisation set) of proteins
  - classify sites (18 types)
  - collect probability of each residue type in each site type

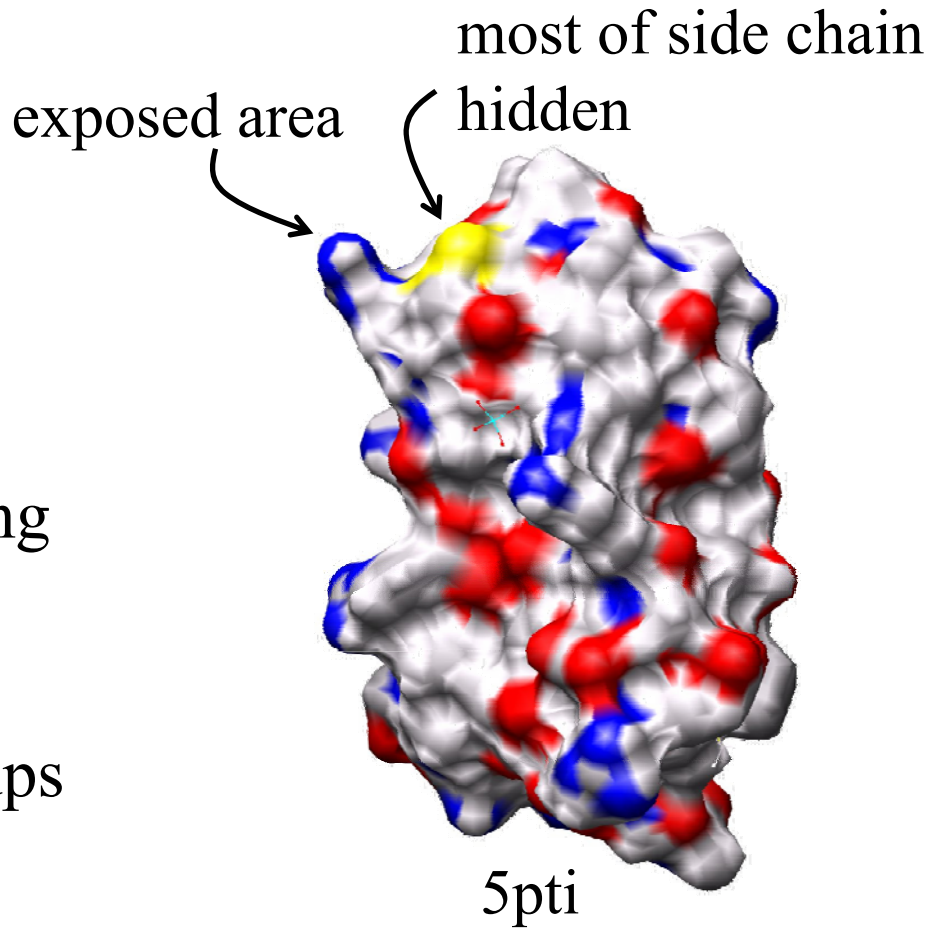
Bowie, J.U., Lüthy, R, Eisenberg, D. (1991) Science 253, 164-170

Environment description

- for each site measure the Å^2 exposed to solvent

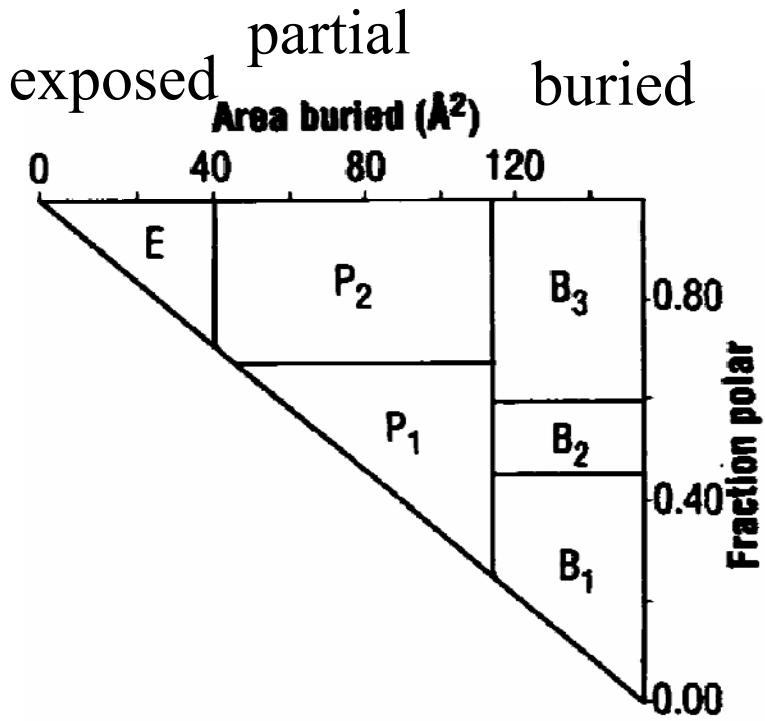
- maybe sometimes one has charges / polar groups touching others
  - measure fraction of buried area covered by polar groups

- define environments…



Environment description

- 6 environment types
- 3 secondary structure types
  - α, β, others
- = 18 environments



- data collection
  - 16 proteins
  - find environment of each site
  - count
    - how many times does one see residue type *i* in environment *j* = *N*(*i*,*j*)
  - count – how many times does one see residue type *i* = *N*(*i*)

Bowie, J.U., Lüthy, R, Eisenberg, D. (1991) Science 253, 164-170

Environment description

- how unusual is a residue *i* in environment *j* ?

$$score(i,j) = \ln\left(\frac{N(i,j)}{N(i)}\right)$$

- final result ? a big scoring table

unlikely

likely

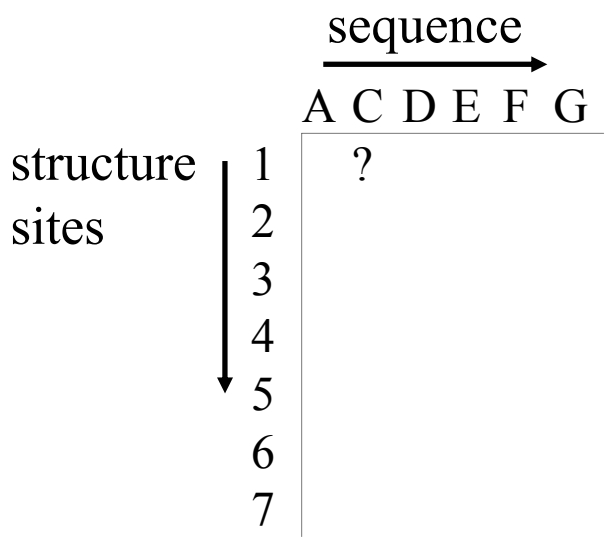
what one expects

Environment class	W	F	Y	L	I	V	M	A	G	P	C	T	S	Q	N	E	D	H	K	R
$B_1\ \alpha$	1.00	1.32	0.18	1.27	1.17	0.66	1.26	-0.66	-2.53	-1.16	-0.73	-1.29	-2.73	-1.08	-1.93	-1.74	-1.97	-0.34	-1.82	-1.67
$B_1\ \beta$	1.17	0.85	0.07	1.13	1.47	1.09	0.55	-0.79	-2.02	-0.94	-0.22	-1.12	-2.91	-1.67	-1.42	-1.93	-2.56	-1.91	-2.69	-1.16
B_1	1.05	1.45	0.17	1.10	1.11	1.02	0.98	-0.91	-1.92	0.26	-1.22	-1.53	-2.81	-1.17	-2.42	-2.52	-1.76	-1.12	-2.59	-2.16
$B_2\ \alpha$	0.50	0.90	0.85	1.01	0.63	0.68	1.12	-0.69	-1.49	-2.21	-0.10	-1.50	-1.47	-0.23	-0.61	-0.71	-1.62	0.23	-0.78	0.06
$B_2\ \beta$	0.01	1.18	1.06	0.76	1.31	1.06	0.64	-1.55	-2.26	-0.49	-0.87	-2.27	-1.77	-1.22	-2.07	-1.07	-1.41	-0.77	-1.14	-0.20
B_2	1.02	1.05	1.12	0.84	0.81	0.60	0.90	-0.66	-1.66	0.19	-0.05	-0.76	-1.17	-0.76	-0.66	-1.35	-1.28	0.46	-2.34	-0.80
$B_3\ \alpha$	0.92	-0.03	0.58	0.15	0.04	-0.02	0.89	-0.57	-1.86	-0.68	-1.56	-0.57	-0.96	0.22	-0.06	0.08	-0.50	0.73	0.43	0.96
$B_3\ \beta$	0.75	0.81	1.30	0.18	0.54	0.56	-0.57	-0.93	-1.93	-0.34	-0.54	-0.44	-0.74	0.21	-0.24	-0.14	-0.86	0.82	-0.53	0.13
B_3	1.07	0.70	1.13	0.35	-0.17	-0.03	0.23	-0.96	-0.98	-0.13	-1.20	-0.53	-0.54	0.05	0.04	-0.36	-1.05	1.01	0.10	0.66
$P_1\ \alpha$	-1.35	-0.82	-0.59	-0.52	-0.24	0.10	-0.03	0.73	-0.49	-0.25	0.95	0.31	0.34	-0.14	-0.54	-0.17	-0.25	-0.52	-0.21	-0.28
$P_1\ \beta$	0.36	-0.49	0.17	-1.03	0.20	0.46	-0.27	0.64	-0.82	-0.55	1.49	0.93	0.33	-2.27	-1.32	-0.73	-1.07	-0.42	-1.21	-0.77
P_1	-1.26	-1.20	-1.31	-0.62	-0.23	-0.01	-1.19	0.46	-0.24	0.66	1.35	0.56	0.49	-0.63	-0.13	-0.61	0.38	-1.12	-0.74	-1.29
$P_2\ \alpha$	-1.14	-1.43	-0.79	-0.35	-0.54	-0.48	-0.45	0.06	-0.50	-0.26	-0.93	-0.05	-0.18	0.55	-0.05	0.56	0.28	0.06	0.61	0.50
$P_2\ \beta$	-0.79	-0.54	-0.84	-1.30	-0.33	0.13	-0.72	-0.55	-0.98	-1.29	-0.57	0.84	0.59	-0.08	-0.16	0.32	0.19	-0.87	0.59	0.10
P_2	-0.82	-0.86	-0.51	-0.70	-1.09	-0.88	-0.89	-0.15	-0.40	0.44	-0.60	0.06	0.26	0.27	0.50	0.27	0.49	0.13	0.44	0.30
$E\ \alpha$	-1.35	-2.20	-2.10	-1.58	-2.76	-1.10	-0.72	0.46	0.68	0.04	-0.44	-0.17	0.15	0.36	0.28	0.59	0.44	-0.19	0.13	-0.34
$E\ \beta$	0.64	-0.90	0.30	-1.66	-1.47	-1.74	-0.68	0.06	1.46	-0.96	-0.24	0.14	0.65	-0.19	-0.06	-0.16	-0.78	-0.83	-0.52	-0.49
E	-2.14	-1.90	-0.94	-1.19	-1.61	-0.91	-1.67	0.12	1.13	0.20	-0.46	0.12	0.32	-0.03	0.41	0.03	0.22	-0.25	-0.14	-0.32

Bowie, J.U., Lüthy, R, Eisenberg, D. (1991) Science 253, 164-170

Environment description - application

- given these descriptions – use them
- take a protein structure label each site
- take sequence of interest
- for each residue
  - score at each site of protein
- score matrix
- find best path
  - sequence to structure alignment

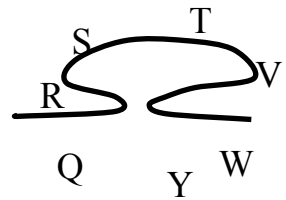


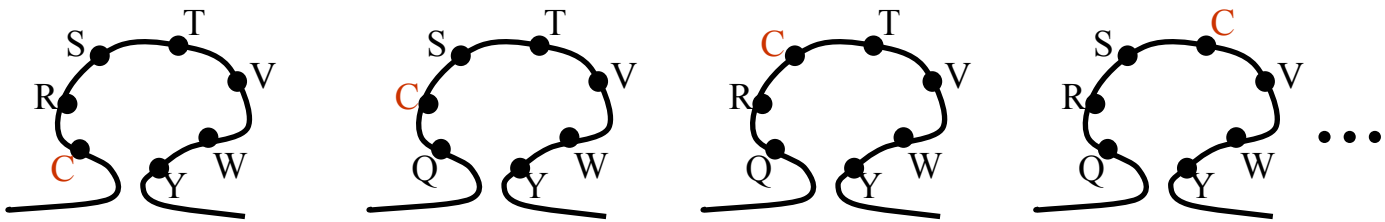
- final application
  - take protein databank
  - try to align your sequence to every structure

Frozen approximation

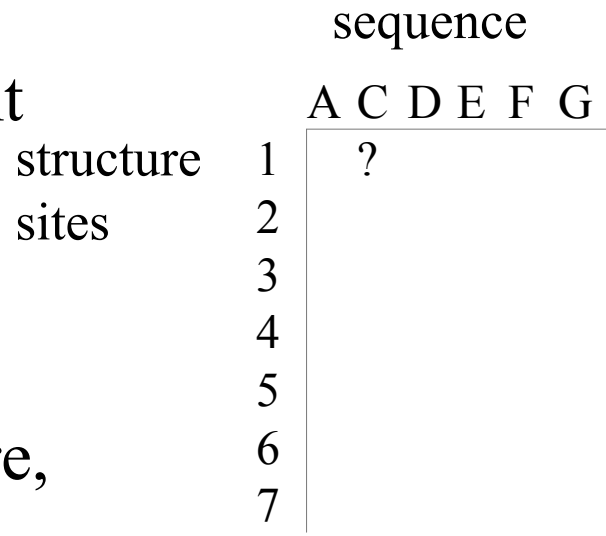
- original problem
  - we want to use a score function which
    - sensitive to sequence
    - sensitive to structure

- remember – original structure did have a sequence
- belief
  - if two proteins are related, the sequences will have similar properties
  - score with the residues of the original sequence





Frozen approximation



- I can score my sequence in the environment of the known structure
- good
  - the environment is well characterised
    - if my structure has polar residues here, they will go into the scoring function
- bad ?
  - we use the sequence of template (known structure)
  - it may only allow very related residues
  - original aim was to move away from close sequences

Summary so far

- look for closely related templates
- try sequence based methods
- sequence to structure methods are definitely possible

- can I make better scoring schemes ?

Scoring schemes

- how much structural information is hidden in sequence ?
- look at a sequence … S T D G W Y F I L S T …
  polar / charged | small | hydrophobic | polar
- I already have labels for sites
  - implicit in substitution matrices
- problem
  - there are lots of exceptions
    - think of mutants – occasional unusual residue does not kill you
  - how to remove the exceptions ?
    - use sequence profiles (psi-blast)
- does the structure contain extra information ? …

Extra information from structures

Residues exist in a protein for different reasons

- gly is easy to substitute – look at diagonal in blosum matrix
- in some turns, gly is essential
  - can only be seen from structure
- cys
  - sometimes a normal hydrophic residue
  - sometimes the geometry says it must form a disulfide bond
  - structure can say if there is another cys near in space
- …

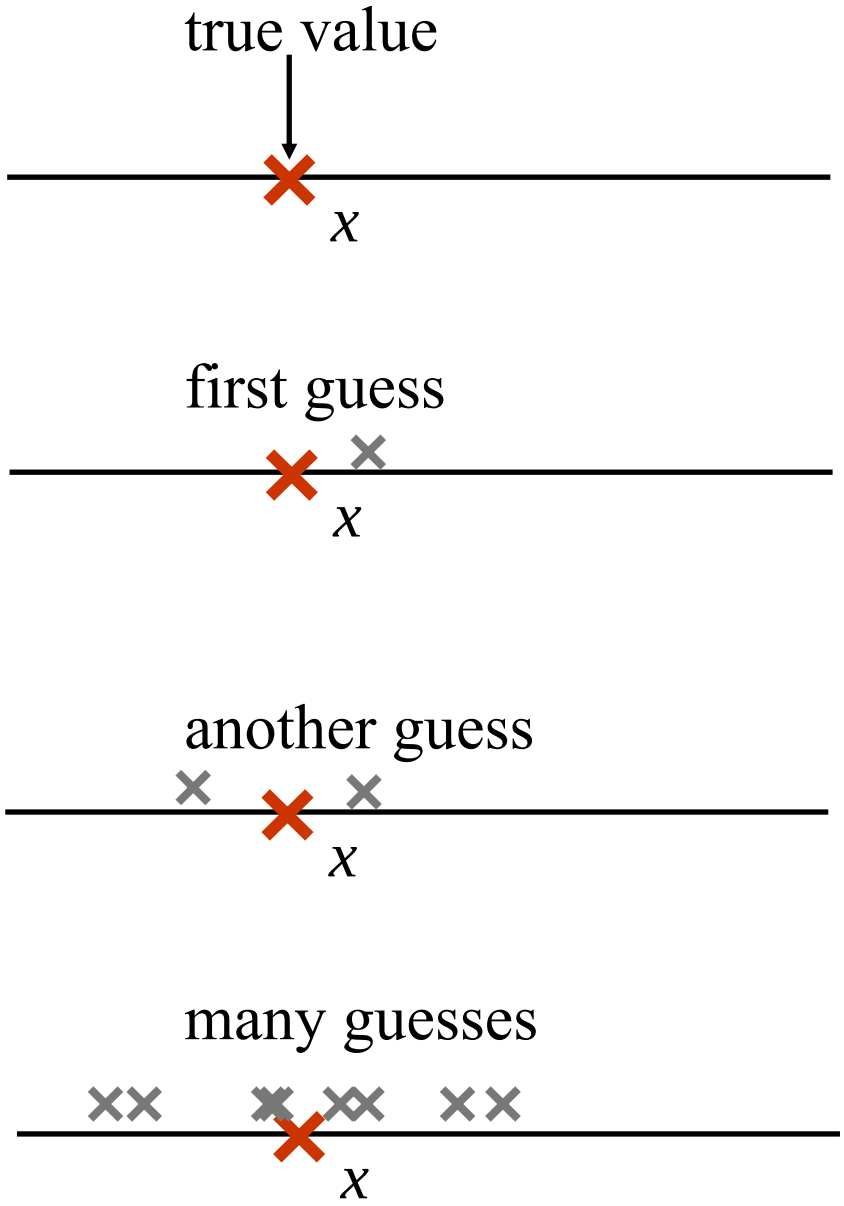
- it should be useful to combine sequence and structure information

Extra information from structures

- Noise arguments

- I have a quantity to estimate
  - measurements contain noise
- as I do more measurements
  - my average estimate is more likely to be correct

- requirement ?
  - the error is not systematic
  - the errors are independent



Extra information from structures

- Claim – hope
  - sequence information (evolution) has statistical noise
  - structure-based methods have noise
  - combination of methods has better signal / noise
- implementation ? easy in principle

  - for each residue *i* in your query sequence
    - for each site *j* in template
      - calculate sequence score s_1 based on profile of *i*
      - calculate structural score s_2 based on fitting residue type *i* into site *j*
      - score for alignment matrix = $s_1 + k\, s_2$
  - for some constant *k*

In practice

- most fold recognition programs combine sequence terms and structural scores
- results may or may not be better than best pure sequence methods

- problems
  - ranking of guesses
  - confidence

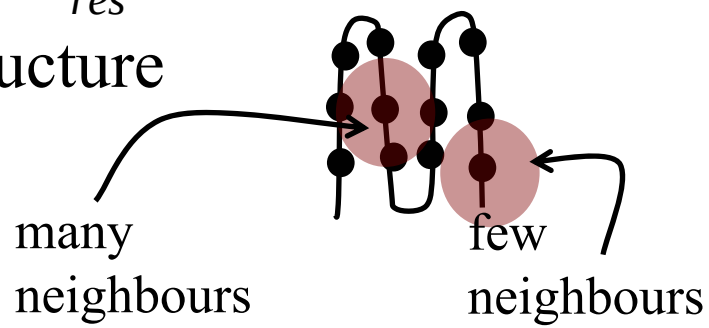
scoring sequence-structure alignments

- I have a sequence which does not fit to a structure
- I have a very good alignment method
  - finds the best arrangement of residues on wrong template
  - they may score well
- may be difficult to tell this from the correct answer

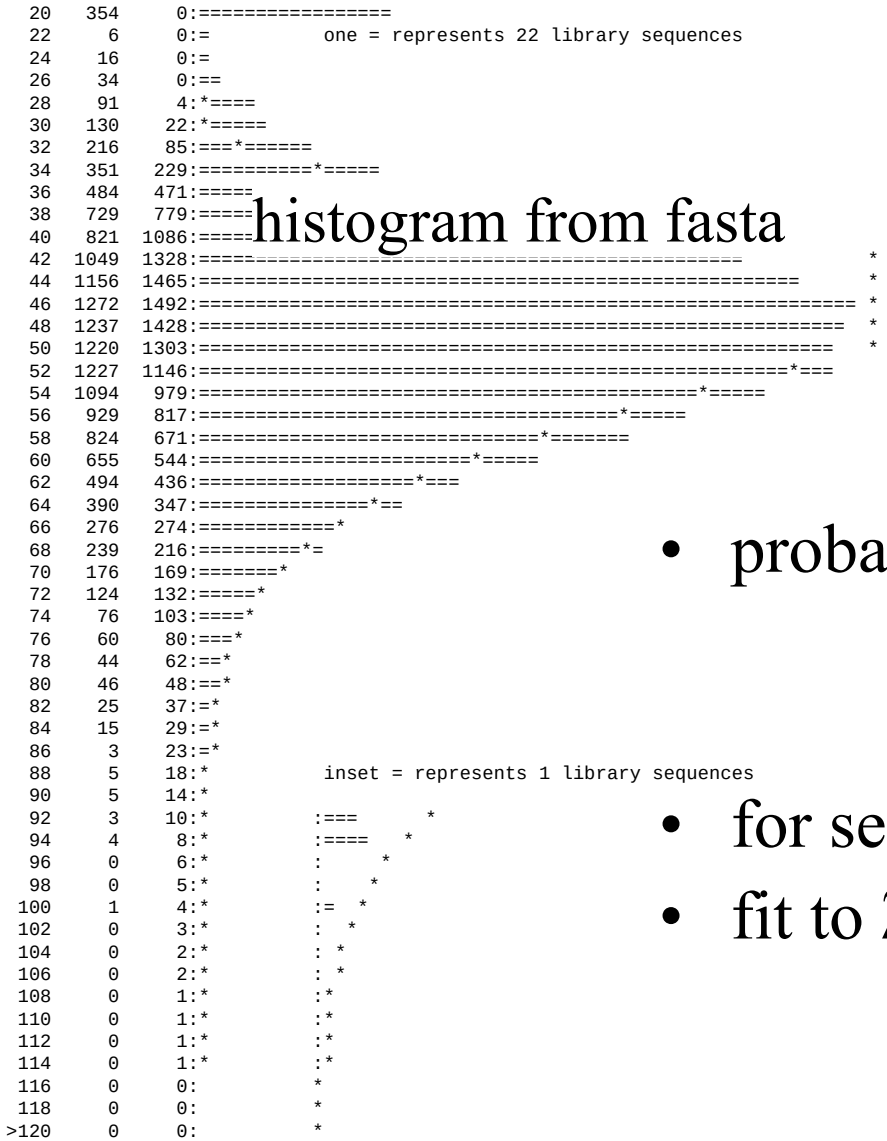
- why are my sequence-structure alignment scores not reliable ?

- score due to sequence approximately $\propto N_{res}$
- score due to structure ? depends on structure

- difficult to judge "good" score
  - different to sequence case



Sequence statistics



- probability P(*S*)of a score greater than *S*

$$P(S) = 1 - \exp\left(-kmne^{-\lambda S}\right)$$

- for sequences length *m*, *n*
- fit to λ and *k* for each sequence

Statistics -problems

- statistics as from sequence alignment assume that score grows with alignment length in a predictable manner
- not the case with structure scores

- real cases
  - not possible to say which remote homology method is best

Summarise and stop

- Use sequence information when possible
- use adventurous sequence methods when necessary
- use very speculative methods (sequence to structure) when necessary