Protein Fold Recognition / Weak Similarities

Why do we do sequence alignments ?

- find related proteins
  - build models
  - guess at function

For some interesting protein

- sequence always available

What should one do with really weak sequence homology ?

Two ideas

- how to search for very weak similarities
- can one take advantage of conserved structures ?

Technical

- Searching for remote sequence homologues

- Sequence to structure alignments

Assumed knowledge

- Some memory of sequence alignment methods, score matrix, $\mathrm{O}(n^2)$ cost

Mission

For some protein sequence – find as much as possible

- function
- build good model
- build a bad model

Vague information may be useful

- which residues are near active site ?
- which residues are near a dimer interface ?
- which residues are in weakly structured loops ? (chemical modification)
- bad model may be enough for phasing (X-ray)

Approach

- start with most reliable methods
- add more speculative methods as necessary

Example

- simple sequence searches
- searches for more remote homologues
- searches for possible structures

Methods in other courses

- emphasis on speed (in Georgio's lectures)

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 alignment	yes	no
very remote homologues	may work	less likely to work

Does speed matter ?

Slow methods

Methods for large databases are

- fast
- approximate

Here

- ultimate use is often a small database (PDB 9.7 × 10^4)
- computer time does not matter

In lab you have 1 or 10's of proteins

- each take weeks or months to work on
- if each search takes hours ? no problem

Remote searches…

Remote searches

When to do this ?

Assume simple (blast / fasta) search returned

- related sequences
- unknown function
- none of related proteins have known structures

Weak sequence similarities

Your sequence

A B D E F G H I K L M N P Q...

finds no helpful proteins. Try searching with a related protein

prot_1 **A B Q E F G R I S L T N P Q...**

- finds a protein whose structure has been solved

prot_2 **Q B Q E Q G R Q S L T N P A...**

Claim

- yours & prot_2 are related
- relationship too weak to see directly
- prot_2 can be used
    - to make a bad model, guess for function

Weak sequence similarities

- first idea
- take your protein
- collect related proteins
    - foreach (related protein)
        - do a sequence search
        - see if results change

- not practical
- not very systematic

- what else does one get from homologues ?

Information from related sequences

- usually one finds many related sequences.

- consider details…

```
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGAHAGEYGAEALEKMFLSFPTTKTYFPHFDLSHGSAQVKGHG
 LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGAHAGDYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPDDKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTHVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGAHAGEYGAEAWERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAAWGKVGAHAGEYGAEAWERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAYWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAHWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSAADKTNVKAGWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG
VLSAADKTNVKAFWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG
VLSADDKANIKAEWGKIGGHGAEYGAEALERMFCSFPTTKTYFPHFDVSHGSAQVKGHG
MLSPADKTNVKADWGKVGAHAGEYGAEAFERMFLSFPTTKTYFPHFDLSHGSAQVKGQG
VLSPADKTNVKACWGKVGAHAGEYGAEAFERMFLSFPTTKTYFPHFDLSHGSAQVKGQA
VLSAADKSNVKAAWGKVGGNAGAYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKSNVKATWDKIGSHAGEYGGEALERTFASFPTTKTYFPHFDLSPGSAQVKAHG
VLSPADKSNVKAWWGKVGGHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
MLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTGTYFPHFDLSHGSAQVKGHG
VLSSADKNNVKACWGKIGSHAGEYGAEALERTFCSFPTTKTYFPHFDLSHGSAQVQAHG
VLSAADKSNVKAAWGKVGGNAGAYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSPADKTNVKAQWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
VLSANDKSNVKAAWGKVGNHAPEYGAEALERMFLSFPTTKTYFPHFDLSHGSSQVKAHG
VLSPADKSNVKAAWGKVGGHAGDYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
    …       …       … …
```

Conservation

If your sequence has a Q here,

- may not be helpful to use it in sequence searches

```
L D D Q R Q   S T R
L D A Q R A D S T R
V D D Q R R W S T R
A D D Q R C A S S K
I D D Q R D D S T R
L D D Q R E G S T K
L D D Q R F C S T R
```

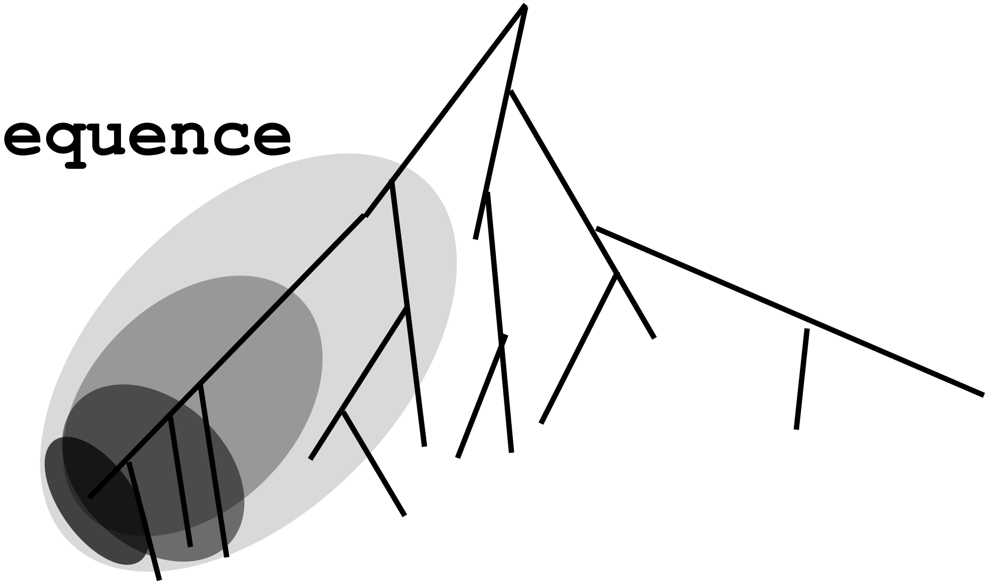
- better to use the "average" residue at this point

- first have to find the "average" residue
- leads to method

Searching with profiles

- initial average_sequence = your_sequence

```
while (step < max_steps)
   search with blast using average_sequence
   if interesting result (function /
   structure..)
      return results
   else
      update average_sequence
```



- basis of "psi-blast"

- does it work ?

Remote sequence searching

- much more sensitive than simple searches, but

- involves weaker sequence similarities, more errors
- alignment not perfect
- statistical significance harder to estimate
- possibility of finding unrelated sequences (rubbish)

- still relies on some significant sequence similarity

- can one move away from sequence similarity ?

Why move away from sequence ?

- if sequences provide information – use this

- if you are desperate…

Sequence alignments – implied structures

From sequence viewpoint

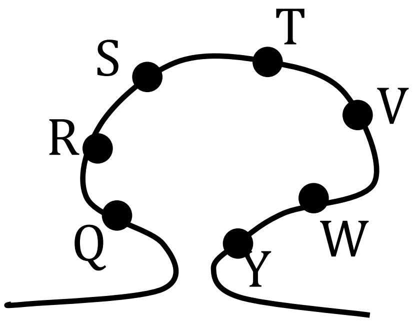
..**AC-DEFG**..

..QRSTVWY..

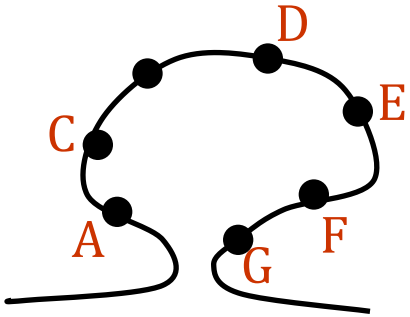
What if structure of second sequence is known ?

..**AC-DEFG**.. query sequence

..**QRSTVWY**.. known structure



known
structure



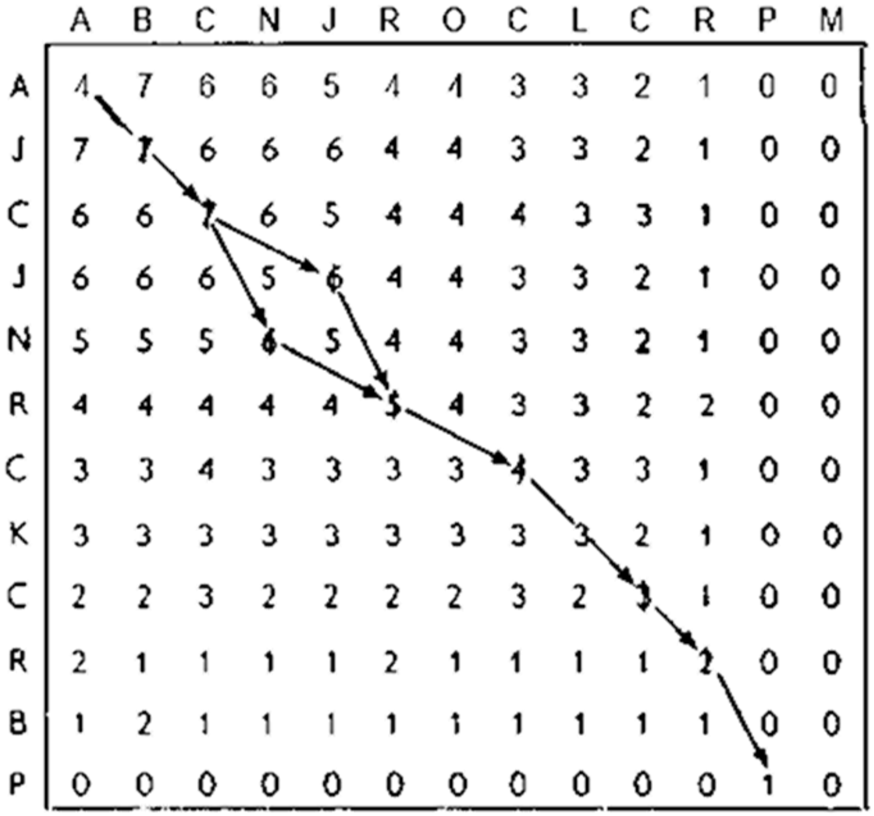
model
implied

Sequence to structure alignments

Remember how sequence alignments work

- similarity / substitution scores
- fill out score matrix
- find best path

Can we use this for sequence to structure alignments ?



	A	B	C	N	J	R	O	C	L	C	R	P	M
A	4	7	6	6	5	4	4	3	3	2	1	0	0
J	7	7	6	6	6	4	4	3	3	2	1	0	0
C	6	6	7	6	5	4	4	4	3	3	1	0	0
J	6	6	6	5	6	4	4	3	3	2	1	0	0
N	5	5	5	6	5	4	4	3	3	2	1	0	0
R	4	4	4	4	4	5	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	4	3	3	1	0	0
K	3	3	3	3	3	3	3	3	3	2	1	0	0
C	2	2	3	2	2	2	2	3	2	3	1	0	0
R	2	1	1	1	1	2	1	1	1	1	2	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

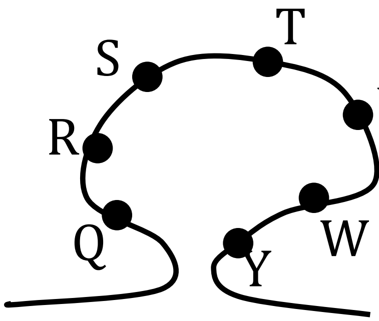
Needleman, SB & Wunsch, CD, (1970) J. Mol. Biol. 48, 443-453

more exotic scoring

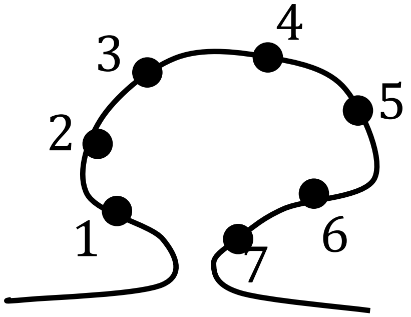
From sequence viewpoint

```
..AC-DEFG..   my sequence
..QRSTVWY..   a protein of known structure
```

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



known structure



forget sequence

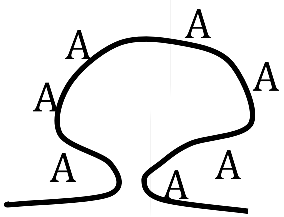
Score matrix ?

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

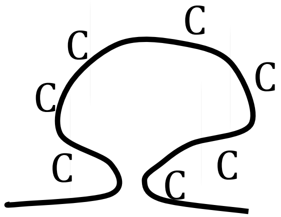
sequence to structure scoring

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

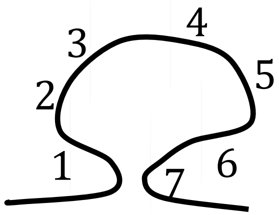
	A	C	D	E	F	G
1	?	...				
2						
3						
4						
5						
6						
7						



then



- then it would be easy to do sequence to structure alignments



Advantage:

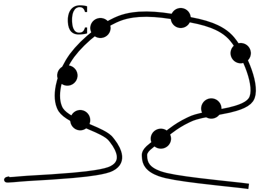
- we claim that structure is more conserved than sequence
- can find appropriate/fitting/suitable structures for a sequence
- very remote, but homologues

vorsicht !!!!

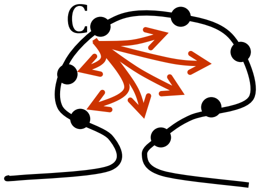
sequence to structure scoring

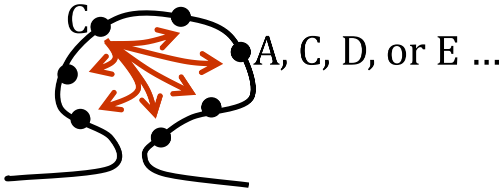
Define an energy function

- depends on interaction of residue with structure
  - easy



- depends on interaction with neighbours
  - but who are the neighbours ?





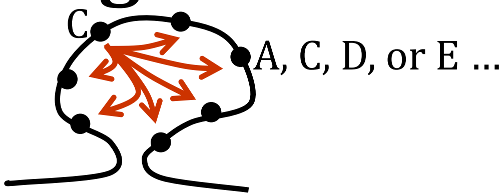
Bad news

- 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

- forget for these lectures

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

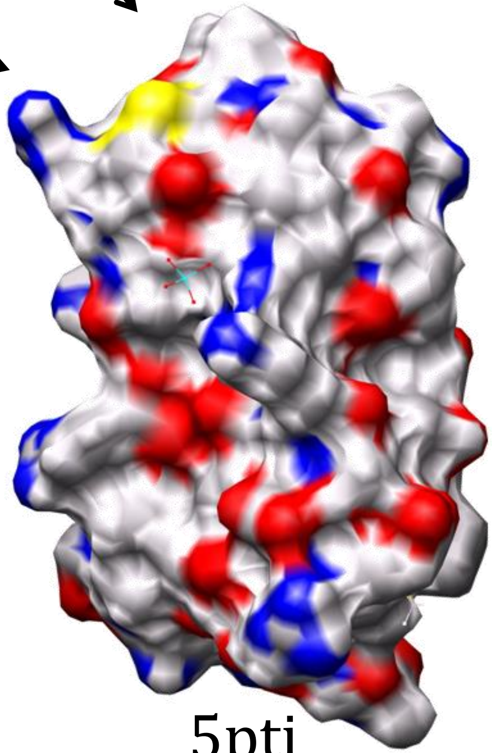
Environment description

For each site measure the Å^2 exposed to solvent

Sometimes one has charges / polar groups touching others

- measure fraction of buried area covered by polar groups

Define environments…



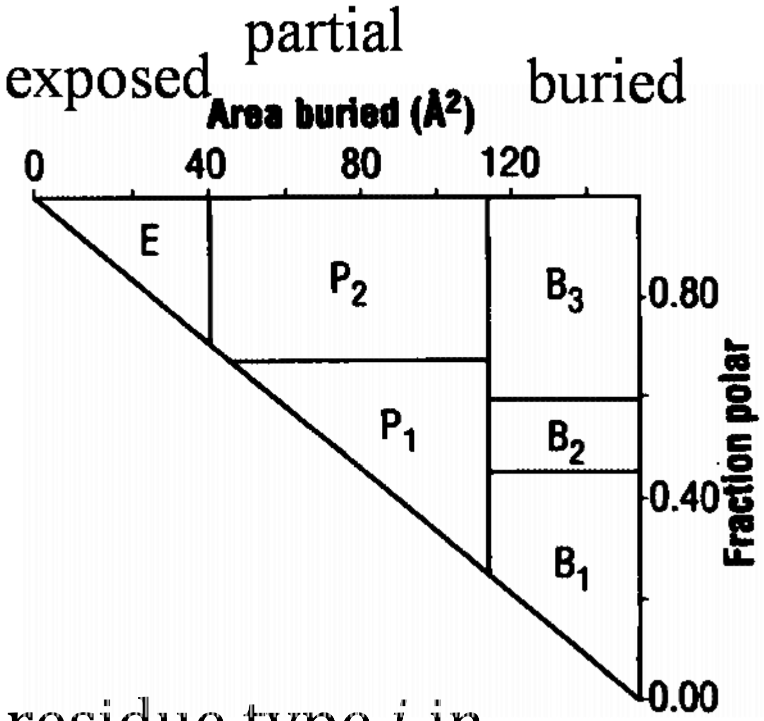
exposed area

most of side chain hidden

5pti

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 often 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* ?

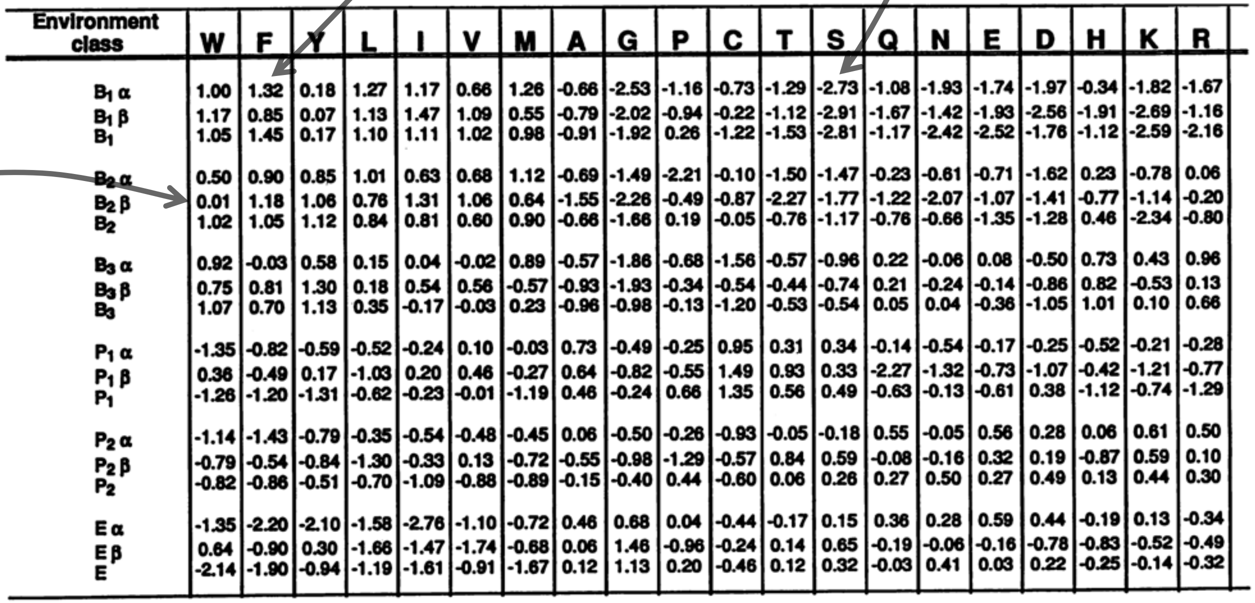
$$score(i,j) = \ln\left(\frac{N(i,j)}{N(i)}\right)$$

Final result ? a big scoring table

unlikely

likely

what one expects

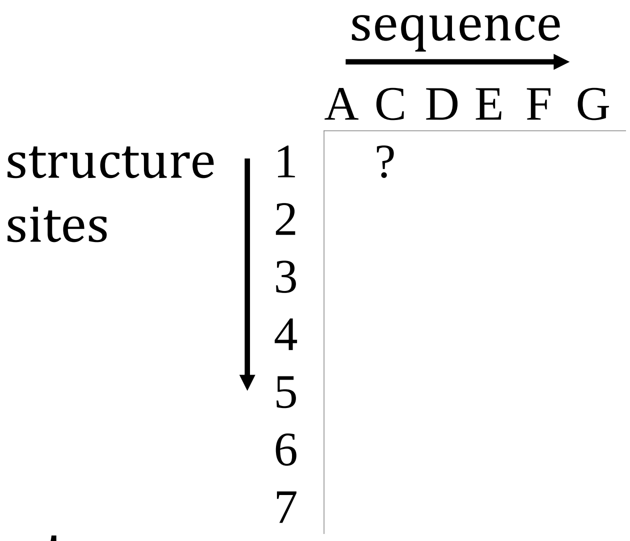


Environment class	W	F	Y	L	I	V	M	A	G	P	C	T	S	Q	N	E	D	H	K	R
$B_1\alpha$	1.00	1.32	0.18	1.27	1.17	0.66	1.26	-0.66	-2.53	-1.16	-0.73	-1.29	-2.73	-1.08	-1.93	-1.74	-1.97	-0.34	-1.82	-1.67
$B_1\beta$	1.17	0.85	0.07	1.13	1.47	1.09	0.55	-0.79	-2.02	-0.94	-0.22	-1.12	-2.91	-1.67	-1.42	-1.93	-2.56	-1.91	-2.69	-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



sequence

structure sites

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

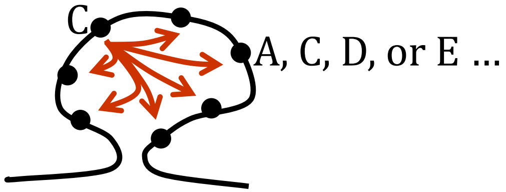
Final application

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

Frozen approximation

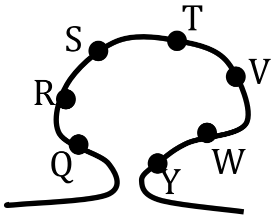
Original problem

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

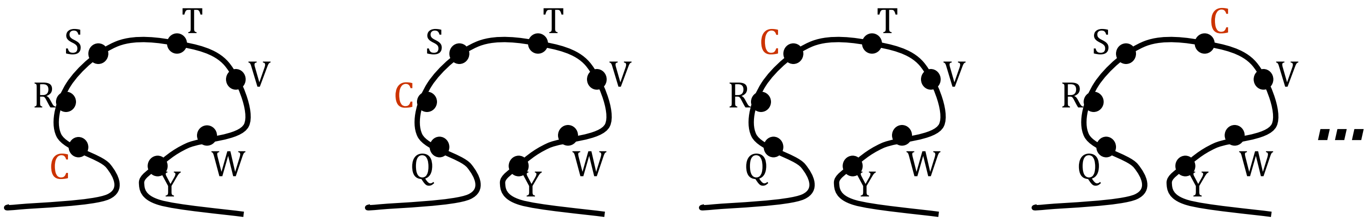


Remember – original structure did have a sequence

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



known structure



Frozen approximation

I can score my sequence in the environment of the known structure

sequence

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

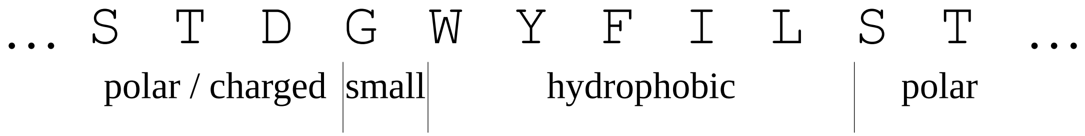
- good
  - the environment is well characterised
    - if my structure has polar residues here, they will go into the scoring function
- bad ?
  - we use the sequence of template (known structure)
  - it may only allow very related residues
  - original aim was to move away from close sequences

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



… S T D G W Y F I L S T …

polar / charged | small | hydrophobic | polar

- how much structural information is hidden in sequence ?
- look at a sequence
- I already have labels for sites
  - implicit in substitution matrices
- 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

Claim – hope

- 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..

Problems with clever methods

Simple sequence searches

- good models for statistical significance
  - (is a related protein really related ?)

Remote sequence searches (psi-blast)

- statistics OK, but less reliable

Structure / Sequence+structure methods ?

- no good model for scores
- no good model for statistical significance

how will score grow with

- size of query ?
- size of alignment ?
- sequence composition ?

Principle

If you have extra information (structure)

- must be a good idea to use it

	sequence	structure based
database size	4×10^7	10^5
	fast	slow
scores	good models	weaker
statistical significance	good or almost good	weaker

Summarise and stop

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