Generation, Visualisation and Analysis of Multiple Sequence Alignments

Geoff Barton
Division of Computational Biology
School of Life Sciences
University of Dundee, UK

twitter:@gjbarton
blog: geoffbarton.wordpress.com

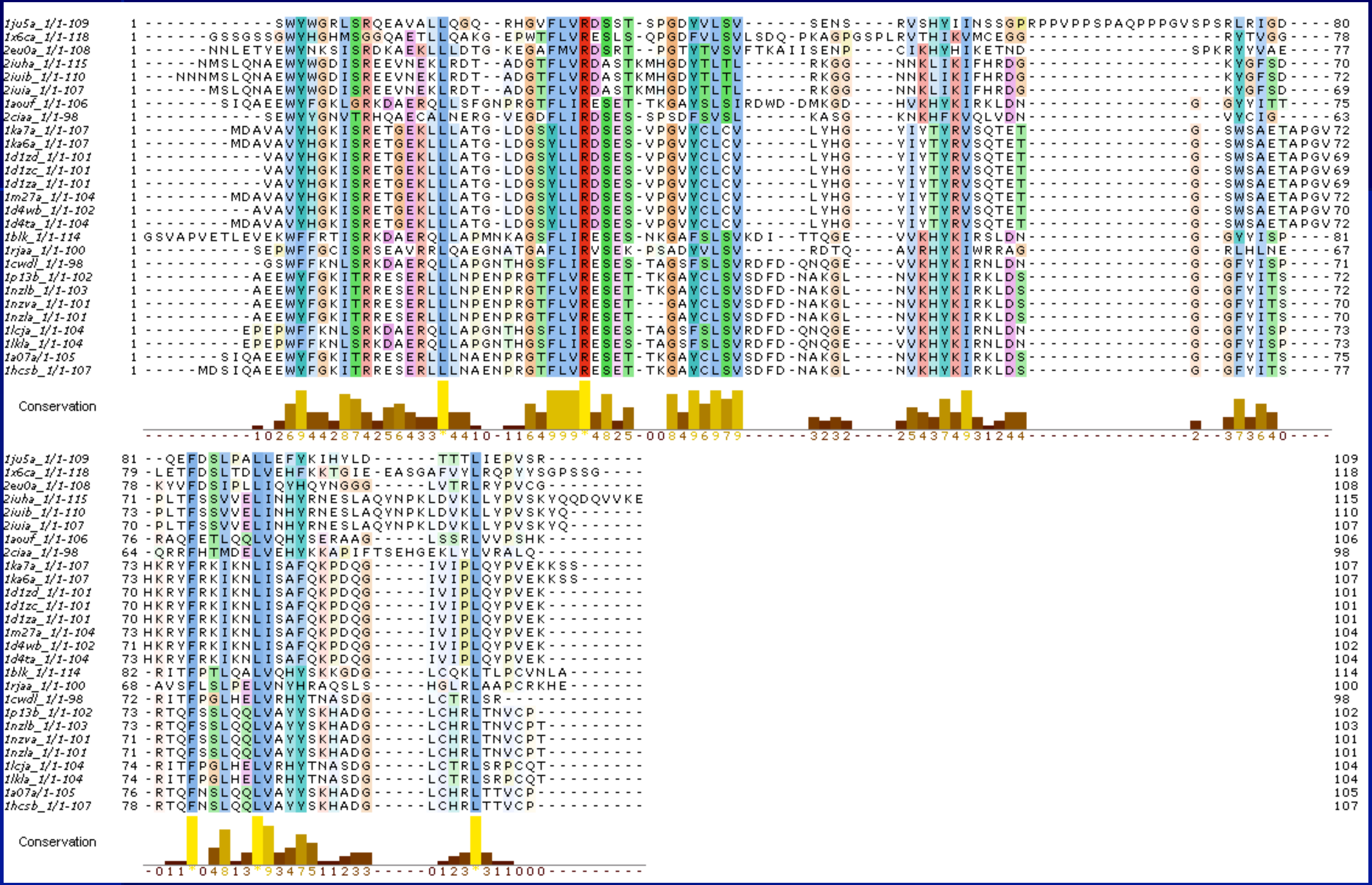
www.compbio.dundee.ac.uk
www.jalview.org

What is multiple sequence alignment (MSA)?

- Alignment of three or more sequences

- What does one look like?

Example Multiple Sequence alignment of 27 SH2 domains



```
1ju5a_1/1-109    1 ------------SWYWGRLSRQEAVALLQGQ--RHGVFLVRDSST-SPGDYVLSV------SENS----RVSHYIINSSGPRPPVPPSPAQPPPGVSPSRLRIGD----80
1x6ca_1/1-118    1 ------GSSGSSGWYHGHMSGGQAETLLQAKG-EPWTFLVRESLS-QPGDFVLSVLSDQ-PKAGPGSPLRVTHIKVMCEGG------------------RYTVGG----78
2eu0a_1/1-108    1 ------NNLETYEWYNKSISRDKAEKLLLDTG-KEGAFMVRDSRT--PGTYTVSVFTKAIISENP----CIKHYHIKETND---------------SPKRYYVAE----77
2iuha_1/1-115    1 -----NMSLQNAEWYWGDISREEVNEKLRDT--ADGTFLVRDASTKMHGDYTLTL------RKGG----NNKLIKIFHRDG------------------KYGFSD----70
2iuib_1/1-110    1 ---NNNMSLQNAEWYWGDISREEVNEKLRDT--ADGTFLVRDASTKMHGDYTLTL------RKGG----NNKLIKIFHRDG------------------KYGFSD----72
2iuia_1/1-107    1 ------MSLQNAEWYWGDISREEVNEKLRDT--ADGTFLVRDASTKMHGDYTLTL------RKGG----NNKLIKIFHRDG------------------KYGFSD----69
1aouf_1/1-106    1 -------SIQAEEWYFGKLGRKDAERQLLSFGNPRGTFLIRESET-TKGAYSLSIRDWD-DMKGD----HVKHYKIRKLDN---------------G--GYYITT----75
2ciaa_1/1-98     1 -----------SEWYYGNVTRHQAECALNERG-VEGDFLIRDSES-SPSDFSVSL------KASG----KNKHFKVQLVDN------------------VYCIG-----63
1ka7a_1/1-107    1 --------MDAVAVYHGKISRETGEKLLLATG-LDGSYLLRDSES-VPGVYCLCV------LYHG----VIYTYRVSQTET---------------G--SWSAETAPGV72
1ka6a_1/1-107    1 --------MDAVAVYHGKISRETGEKLLLATG-LDGSYLLRDSES-VPGVYCLCV------LYHG----VIYTYRVSQTET---------------G--SWSAETAPGV72
1d1zd_1/1-101    1 -----------VAVYHGKISRETGEKLLLATG-LDGSYLLRDSES-VPGVYCLCV------LYHG----VIYTYRVSQTET---------------G--SWSAETAPGV69
1d1zc_1/1-101    1 -----------VAVYHGKISRETGEKLLLATG-LDGSYLLRDSES-VPGVYCLCV------LYHG----VIYTYRVSQTET---------------G--SWSAETAPGV69
1d1za_1/1-101    1 -----------VAVYHGKISRETGEKLLLATG-LDGSYLLRDSES-VPGVYCLCV------LYHG----VIYTYRVSQTET---------------G--SWSAETAPGV69
1m27a_1/1-104    1 --------MDAVAVYHGKISRETGEKLLLATG-LDGSYLLRDSES-VPGVYCLCV------LYHG----VIYTYRVSQTET---------------G--SWSAETAPGV72
1d4wb_1/1-102    1 ----------AVAVYHGKISRETGEKLLLATG-LDGSYLLRDSES-VPGVYCLCV------LYHG----VIYTYRVSQTET---------------G--SWSAETAPGV70
1d4ta_1/1-104    1 --------MDAVAVYHGKISRETGEKLLLATG-LDGSYLLRDSES-VPGVYCLCV------LYHG----VIYTYRVSQTET---------------G--SWSAETAPGV72
1blk_1/1-114     1 GSVAPVETLEVEKWFFRTISRKDAERQLLAPMNKAGSFLIRESES-NKGAFSLSVKDI--TTQGE----VVKHYKIRSLDN---------------G--GYYISP----81
1rjaa_1/1-100    1 ----------SEPWFFGCISRSEAVRRLQAEGNATGAFLIRVSEK-PSADYVLSV------RDTQ----AVRHYKIWRRAG---------------G--RLHLNE----67
1cwdl_1/1-98     1 -----------GSWFFKNLSRKDAERQLLAPGNTHGSFLIRESES-TAGSFSLSVRDFD-QNQGE----VVKHYKIRNLDN---------------G--GFYISP----71
1p13b_1/1-102    1 ----------AEEWYFGKITRRESERLLLNPENPRGTFLVRESET-TKGAYCLSVSDFD-NAKGL----NVKHYKIRKLDS---------------G--GFYITS----72
1nzlb_1/1-103    1 ----------AEEWYFGKITRRESERLLLNPENPRGTFLVRESET-TKGAYCLSVSDFD-NAKGL----NVKHYKIRKLDS---------------G--GFYITS----72
1nzva_1/1-101    1 ----------AEEWYFGKITRRESERLLLNPENPRGTFLVRESET---GAYCLSVSDFD-NAKGL----NVKHYKIRKLDS---------------G--GFYITS----70
1nzla_1/1-101    1 ----------AEEWYFGKITRRESERLLLNPENPRGTFLVRESET---GAYCLSVSDFD-NAKGL----NVKHYKIRKLDS---------------G--GFYITS----70
1lcja_1/1-104    1 ---------EPEPWFFKNLSRKDAERQLLAPGNTHGSFLIRESES-TAGSFSLSVRDFD-QNQGE----VVKHYKIRNLDN---------------G--GFYISP----73
1lkla_1/1-104    1 ---------EPEPWFFKNLSRKDAERQLLAPGNTHGSFLIRESES-TAGSFSLSVRDFD-QNQGE----VVKHYKIRNLDN---------------G--GFYISP----73
1a07a/1-105      1 -------SIQAEEWYFGKITRRESERLLLNAENPRGTFLVRESET-TKGAYCLSVSDFD-NAKGL----NVKHYKIRKLDS---------------G--GFYITS----75
1hcsb_1/1-107    1 -----MDSIQAEEWYFGKITRRESERLLLNAENPRGTFLVRESET-TKGAYCLSVSDFD-NAKGL----NVKHYKIRKLDS---------------G--GFYITS----77

Conservation
                   ----------10269442874256433*4410-1164999*4825-008496979------3232----254374931244---------------2--373640----

1ju5a_1/1-109   81 --QEFDSLPALLEFYKIHYLD------TTTLIEPVSR---------           109
1x6ca_1/1-118   79 -LETFDSLTDLVEHFKKTGIE-EASGAFVYLRQPYYSGPSSG----           118
2eu0a_1/1-108   78 -KYVFDSIPLLIQYHQYNGGG-----LVTRLRYPVCG---------           108
2iuha_1/1-115   71 -PLTFSSVVELINHYRNESLAQYNPKLDVKLLYPVSKYQQDQVVKE           115
2iuib_1/1-110   73 -PLTFSSVVELINHYRNESLAQYNPKLDVKLLYPVSKYQ-------           110
2iuia_1/1-107   70 -PLTFSSVVELINHYRNESLAQYNPKLDVKLLYPVSKYQ-------           107
1aouf_1/1-106   76 -RAQFETLQQLVQHYSERAAG-----LSSRLVVPSHK---------           106
2ciaa_1/1-98    64 -QRRFHTMDELVEHYKKAPIFTSEHGEKLYLVRALQ----------           98
1ka7a_1/1-107   73 HKRYFRKIKNLISAFQKPDQG-----IVIPLQYPVEKKSS------           107
1ka6a_1/1-107   73 HKRYFRKIKNLISAFQKPDQG-----IVIPLQYPVEKKSS------           107
1d1zd_1/1-101   70 HKRYFRKIKNLISAFQKPDQG-----IVIPLQYPVEK---------           101
1d1zc_1/1-101   70 HKRYFRKIKNLISAFQKPDQG-----IVIPLQYPVEK---------           101
1d1za_1/1-101   70 HKRYFRKIKNLISAFQKPDQG-----IVIPLQYPVEK---------           101
1m27a_1/1-104   73 HKRYFRKIKNLISAFQKPDQG-----IVIPLQYPVEK---------           104
1d4wb_1/1-102   71 HKRYFRKIKNLISAFQKPDQG-----IVIPLQYPVEK---------           102
1d4ta_1/1-104   73 HKRYFRKIKNLISAFQKPDQG-----IVIPLQYPVEK---------           104
1blk_1/1-114    82 -RITFPTLQALVQHYSKKGDG-----LCQKLTLPCVNLA-------           114
1rjaa_1/1-100   68 -AVSFLSLPELVNYHRAQSLS-----HGLRLAAPCRKHE-------           100
1cwdl_1/1-98    72 -RITFPGLHELVRHYTNASDG-----LCTRLSR-------------           98
1p13b_1/1-102   73 -RTQFSSLQQLVAYYSKHADG-----LCHRLTNVCP----------           102
1nzlb_1/1-103   73 -RTQFSSLQQLVAYYSKHADG-----LCHRLTNVCPT---------           103
1nzva_1/1-101   71 -RTQFSSLQQLVAYYSKHADG-----LCHRLTNVCPT---------           101
1nzla_1/1-101   71 -RTQFSSLQQLVAYYSKHADG-----LCHRLTNVCPT---------           101
1lcja_1/1-104   74 -RITFPGLHELVRHYTNASDG-----LCTRLSRPCQT---------           104
1lkla_1/1-104   74 -RITFPGLHELVRHYTNASDG-----LCTRLSRPCQT---------           104
1a07a/1-105     76 -RTQFNSLQQLVAYYSKHADG-----LCHRLTTVCP----------           105
1hcsb_1/1-107   78 -RTQFNSLQQLVAYYSKHADG-----LCHRLTTVCP----------           107

Conservation
                   -011*04813*934751123 3-----0123*311000---------
```

Why are Protein (multiple) Sequence Alignments Useful?

- Link proteins at the amino acid level
- Allow identification of conserved features
- Allow prediction of functionally important residues
- Basis for phylogenetic tree construction
- Basis for sensitive profile-based sequence database searching
- Basis for training many methods to predict features from sequence – e.g. secondary structure

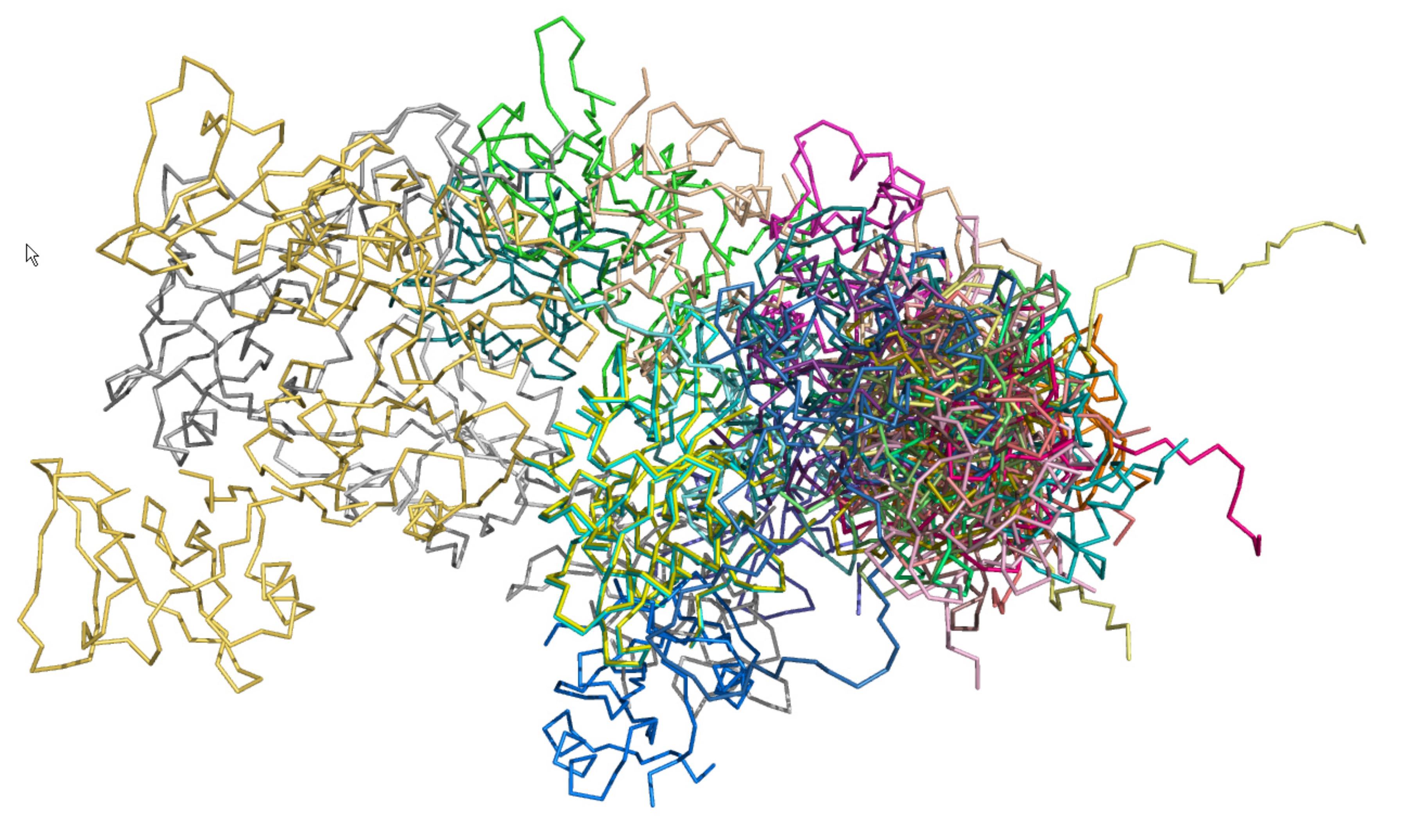
- Standard way of describing and illustrating features of protein sequences and their relationships in publications

Link proteins at the amino acid level

What does this mean?

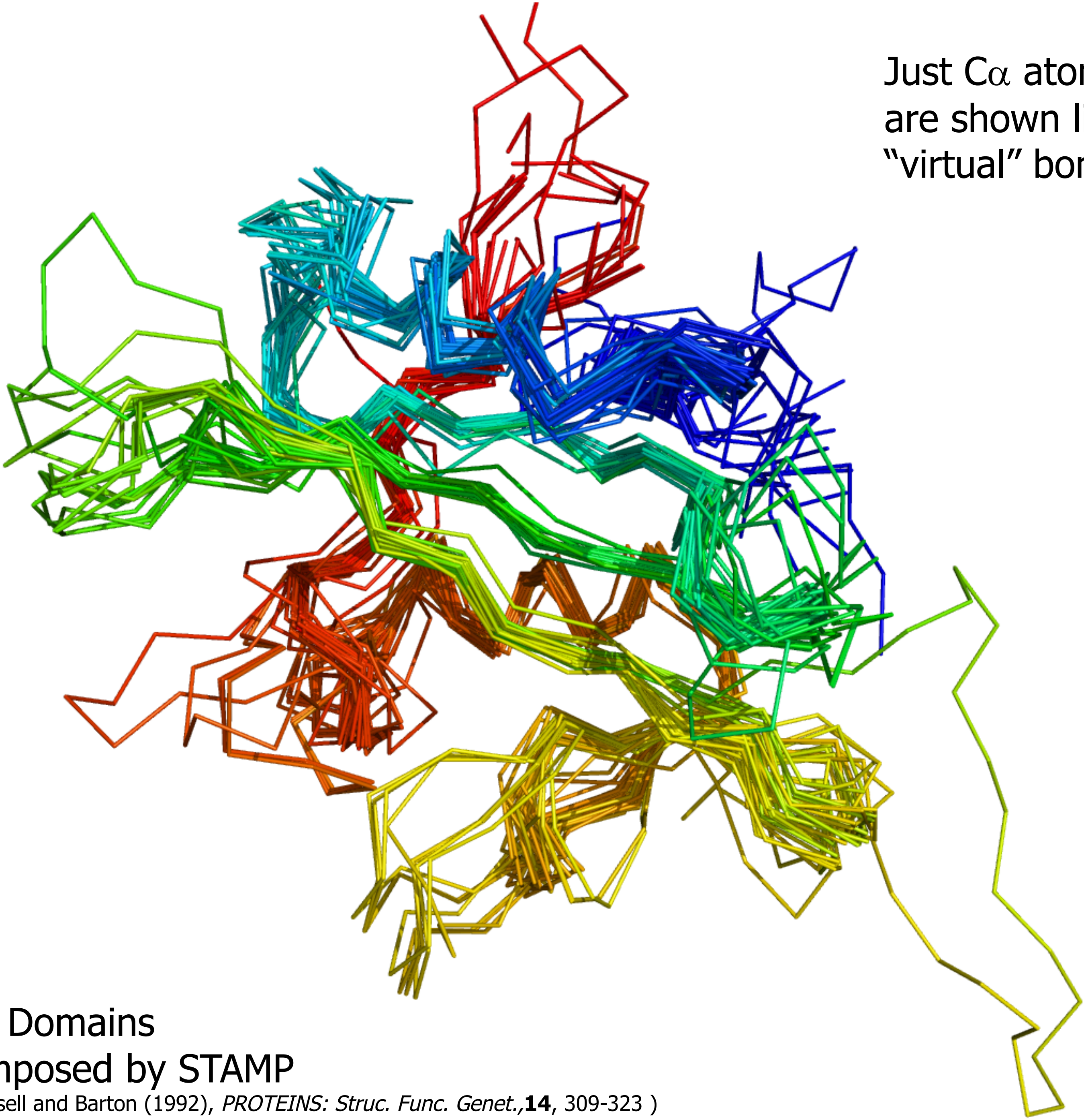
Example sequence alignment of SH2 domains

From the *three-dimensional* structures of the proteins



22 SH2 domain structures as they are if just loaded into PyMol from the PDB

Just Cα atoms are shown linked by "virtual" bonds

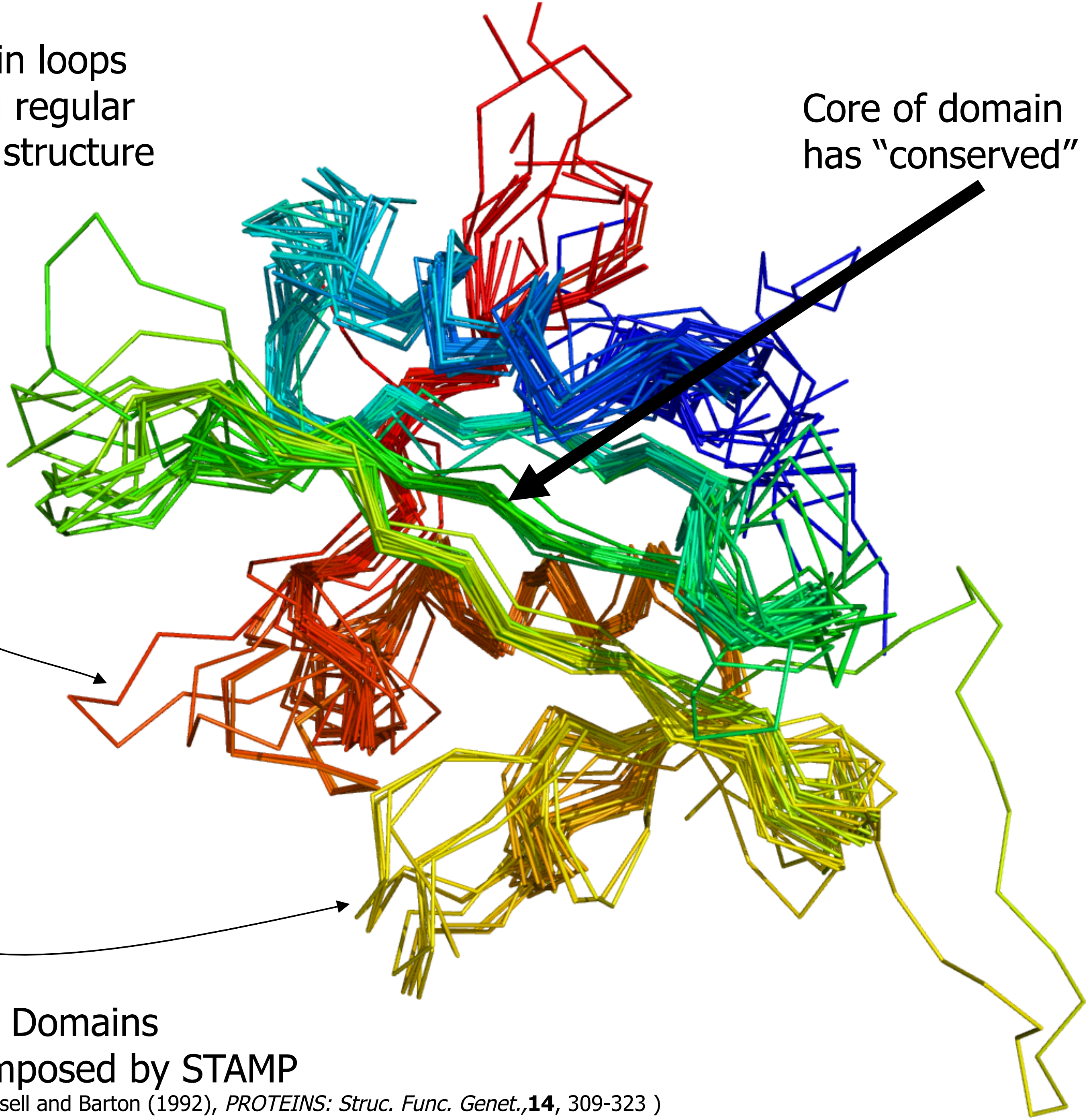


27 SH2 Domains Superimposed by STAMP

(STAMP: Russell and Barton (1992), *PROTEINS: Struc. Func. Genet.*,**14**, 309-323)

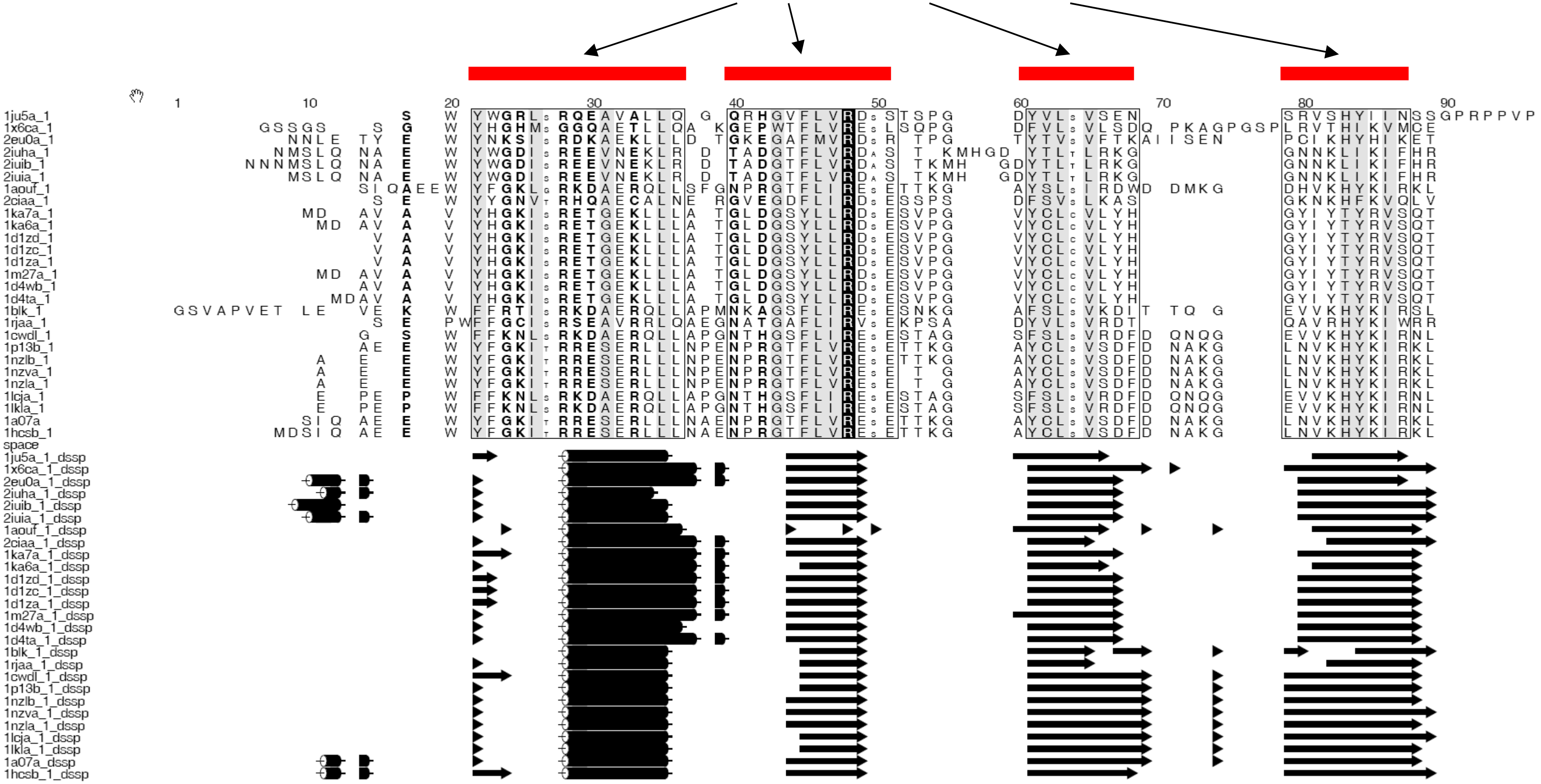
Variability in loops connecting regular secondary structure

Core of domain has “conserved” structure



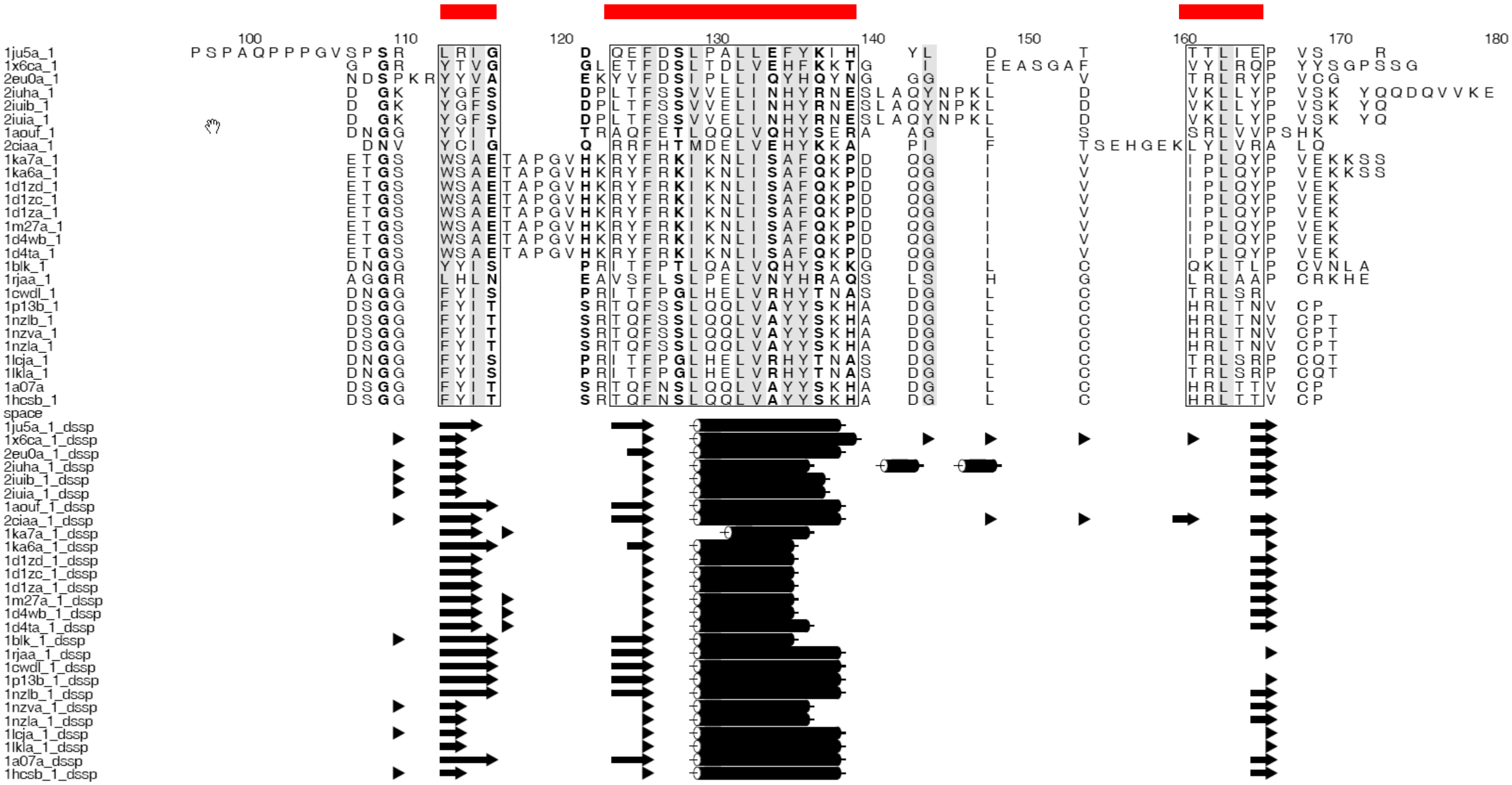
27 SH2 Domains
Superimposed by STAMP
(STAMP: Russell and Barton (1992), *PROTEINS: Struc. Func. Genet.*,**14**, 309-323)

Regions considered by STAMP to be aligned reliably



Structural alignment of 27 SH2 domains showing secondary structure – Part I

STAMP alignment – Alscript display

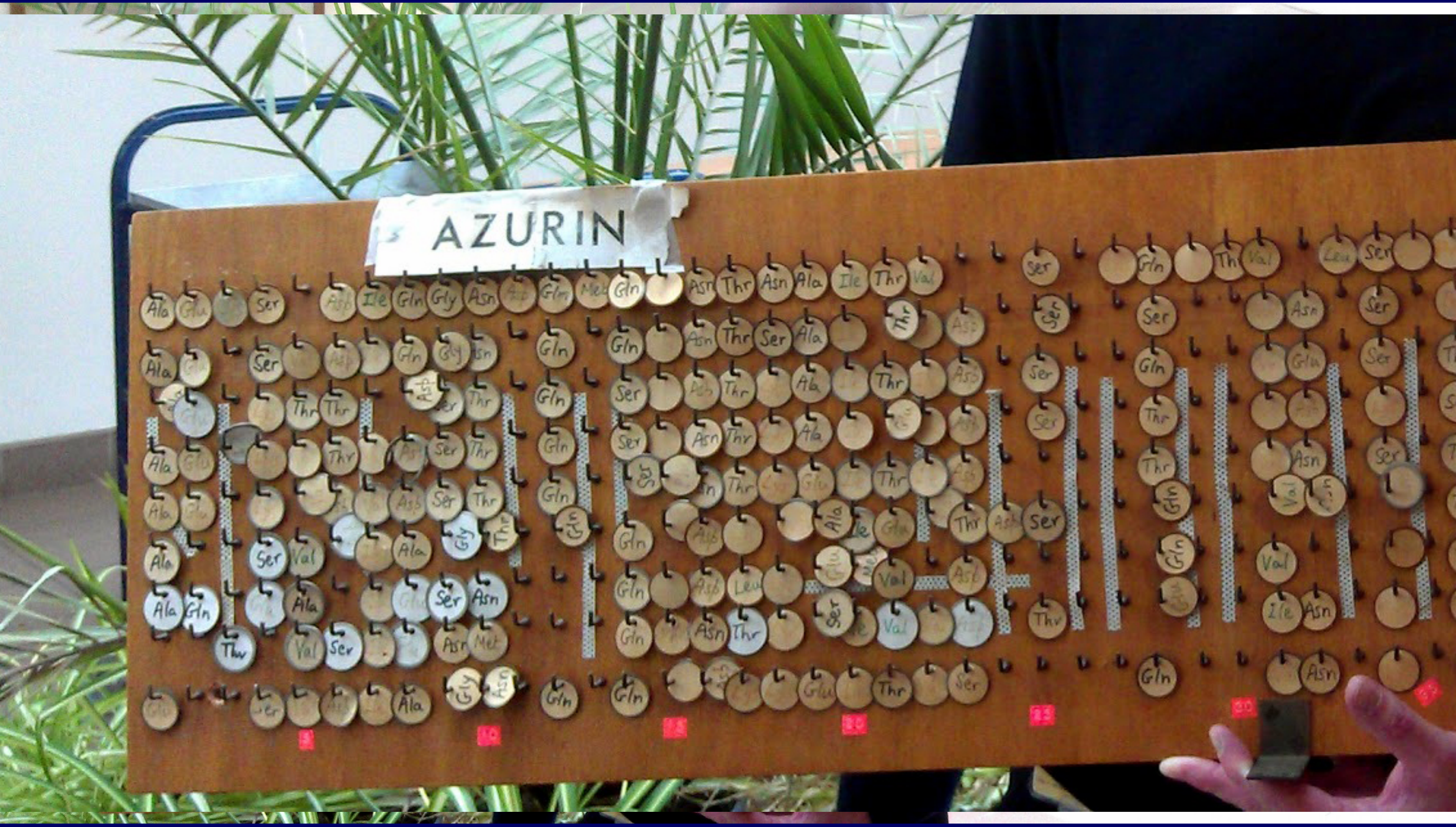


Structural alignment of 27 SH2 domains showing secondary structure – Part II

STAMP alignment – Alscript display

How are MSAs generated when we just have sequences and no knowledge of 3D structure?

Multiple Sequence Alignment from the 1960/70s

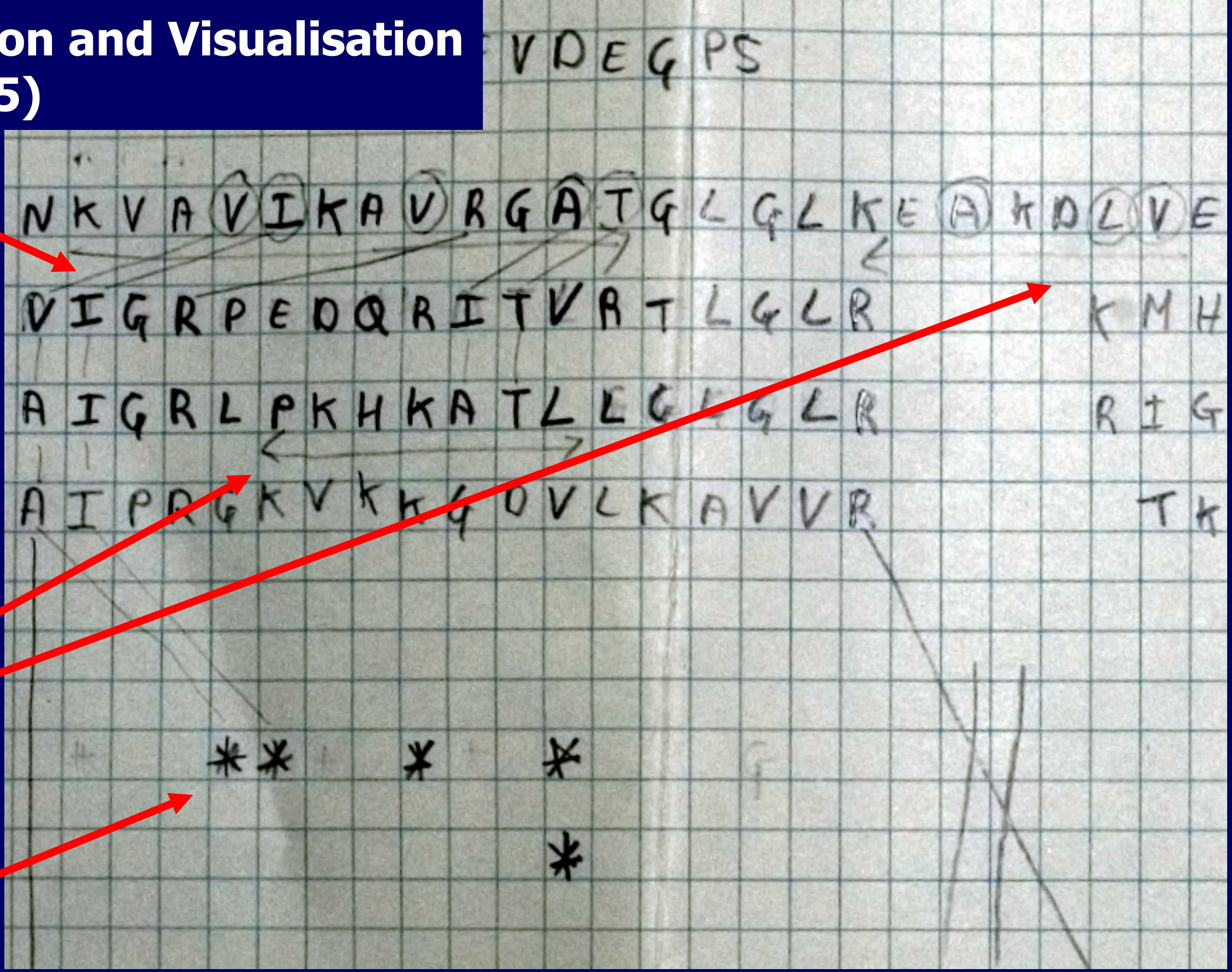


Courtesy of the University of Edinburgh

1984

- Some sequence alignment programs existed but common method was to…
- Align two sequences by writing the amino acid codes on squared paper then sliding them relative to each other to find a good match.
- Use scissors to deal with insertions/deletions
- Yes really!

Multiple Alignment Creation and Visualisation (1984/5)



VDEGPS

NKVAVIKAVRGATGLGLKEAKDLVE

VIGRPEDQRITVRTLGLR KMH

AIGRLPKHKATLLGLGLR RIG

AIPRGKVKKGDVLKAVVR TK

* * * *

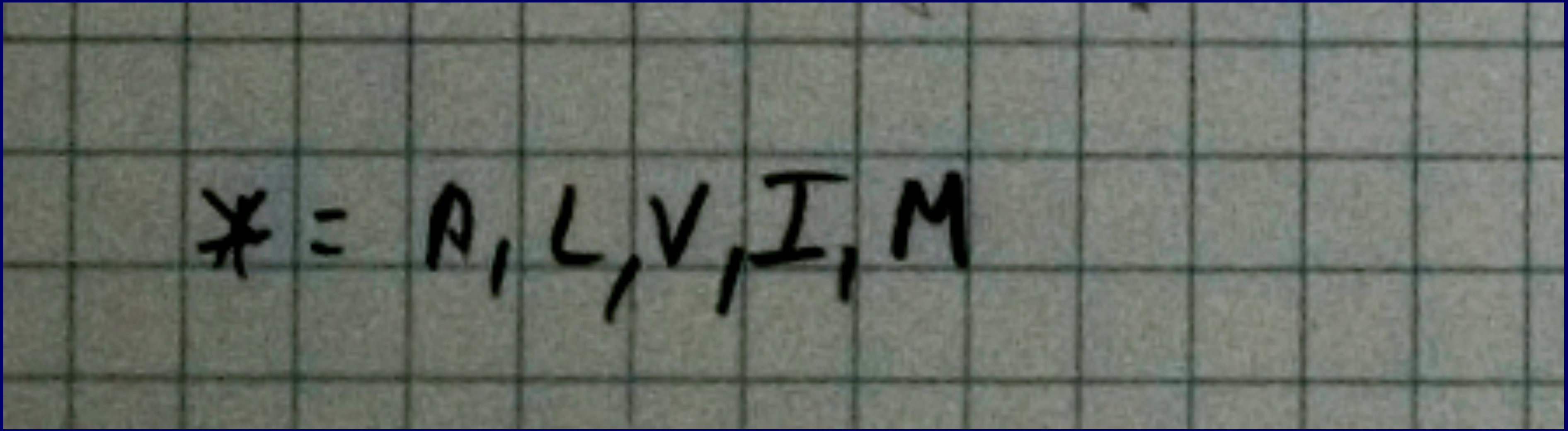
*

"Trace"
(Alternative Alignment)

Location of Beta-Strand
(double headed arrows)

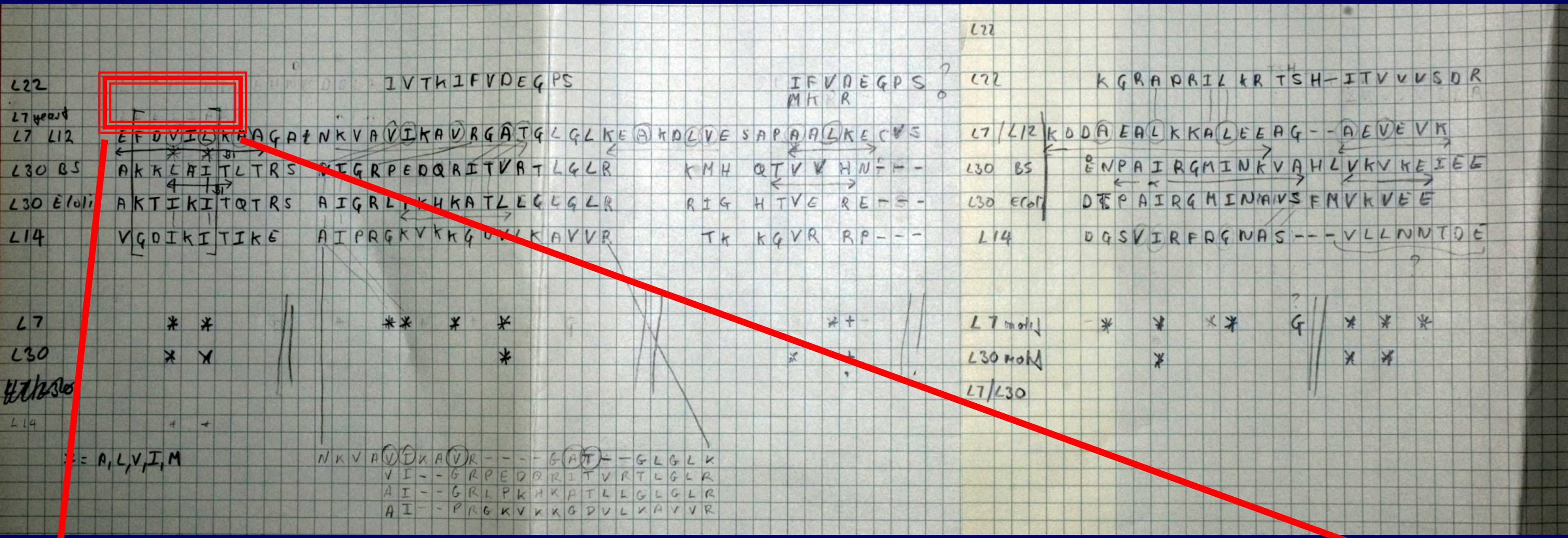
"*" ??

Re-drawn
Alternative Alignment



* = A, L, V, I, M

Multiple Sequence Alignment and Visualisation (1984/5)



L22 IVTKIFVDEGPS IFVDEGPS ?

MHR

L7 yeast

L7 L12 EFDVILKAAGANKVAVIKAVRGATGLGLKEAKDLVESAPAALKECVS

L30 BS AKKLAITLTRS VIGRPEDQRITVRTLGLR KMH QTVV HN---

L30 E.coli AKTIKITQTRS AIGRLPKHKATLLGLGLR RIG HTVE RE-S-

L14 VGDIKITIKE AIPRGKVKKGDVLKAVVR TK KGVR RP---

L7 * * ** * * * +

L30 * * * * +

L14 + +

* = A, L, V, I, M

NKVAVIKAVR----GAT--GLGLK
VI--GRPEDQRITVRTLGLR
AI--GRLPKHKATLLGLGLR
AI--PRGKVKKGDVLKAVVR

L22

L22 KGRAPRILKR TSH-ITVVVSDR

L7/L12 KDDAEALKKALEEAG--AEVEVK

L30 BS ENPAIRGMINKVAHLVKVKEIEE

L30 E.coli DEPAIRGMINIAVSFMVKVEE

L14 DGSVIRFDGNAS---VLLNNTDE

?

L7 motif * * * * ? G * * *

L30 motif * * *

L7/L30

How are MSAs generated?

Pair-wise sequence alignment

Alignment of *two* Protein Sequences -How?

- Need scoring scheme for matching amino acid residues.
- Need to cope with insertions and deletions (gaps or indels).
- Need algorithm to find ‘best’ alignment.
- Need some way of judging if the alignment is likely to be correct.

Protein Scoring Schemes

- A table of scores for aligning each possible amino acid pair.
- **Simplest scheme**, just scores 1 for identity and 0 for non identity.
- Better schemes weight similarities in amino acid properties or observed substitutions. For example, BLOSUM and PAM series. Virtually all of today's programs use these.

BLOSUM62 Matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1	-4
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-4
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

BLOSUM62 is a log-score matrix – more on this later...

Gap Penalties

- Score for aligning a residue or residues in one protein to a gap in the other.
- Most usual form:
  penalty = *ul* + *v*
- where l is the length of the gap and *u* and *v* are constants.
- *u* is often called the gap extension penalty, *v*, the gap creation penalty.

Finding the ‘best’ alignment

- The mathematically best alignment is the one that gives the highest score when the amino acids of the two proteins are aligned, taking account of any gaps.
- *This alignment is not necessarily the one that is biologically meaningful.* (more on this later)

Finding the best alignment

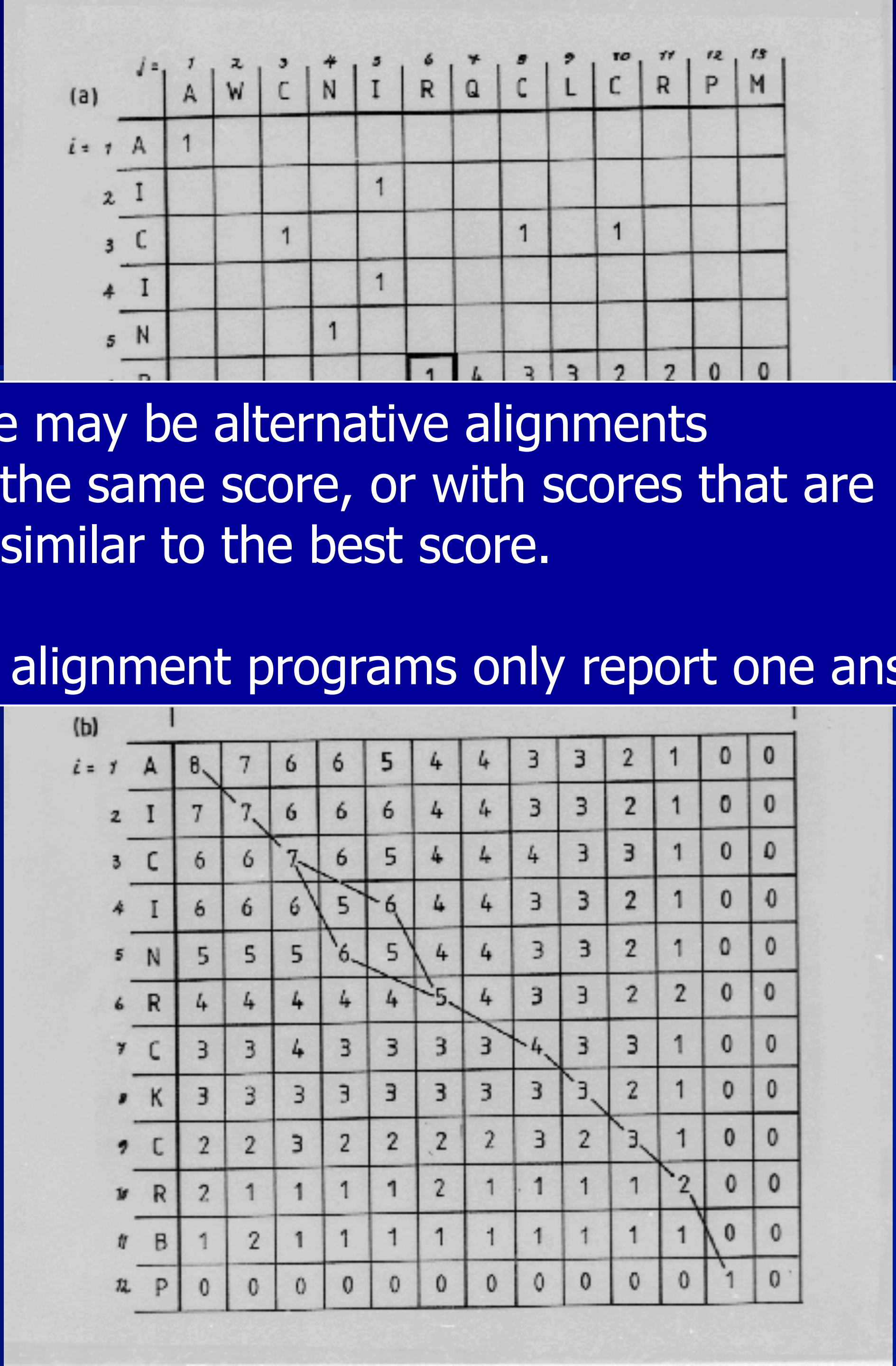
- Naïve way would be to generate all possible alignments of the two sequences, then take the one with the highest score according to the BLOSUM matrix.
- But… for two sequences of 100 amino acids, there are > 10^{75} possible alignments…

Dynamic Programming

- Trick to avoid having to generate all possible alignments.
- First introduced in molecular biology by Needleman and Wunsch (1970).
- Many variations on the theme.
- Basis of (nearly) all sequence alignment programs.
- Finds the mathematically ‘best’ score for alignment of two sequences of length *M* and *N* in *MN* steps.

There may be alternative alignments
with the same score, or with scores that are
very similar to the best score.

Most alignment programs only report one answer…



(a)

j =		1	2	3	4	5	6	7	8	9	10	11	12	13
		A	W	C	N	I	R	Q	C	L	C	R	P	M
i = 1	A	1												
2	I					1								
3	C			1					1		1			
4	I					1								
5	N				1									
	R						1	4	3	3	2	2	0	0

(b)

i = 1	A	8	7	6	6	5	4	4	3	3	2	1	0	0
2	I	7	7	6	6	6	4	4	3	3	2	1	0	0
3	C	6	6	7	6	5	4	4	4	3	3	1	0	0
4	I	6	6	6	5	6	4	4	3	3	2	1	0	0
5	N	5	5	5	6	5	4	4	3	3	2	1	0	0
6	R	4	4	4	4	4	5	4	3	3	2	2	0	0
7	C	3	3	4	3	3	3	3	4	3	3	1	0	0
8	K	3	3	3	3	3	3	3	3	3	2	1	0	0
9	C	2	2	3	2	2	2	2	3	2	3	1	0	0
10	R	2	1	1	1	1	2	1	1	1	1	2	0	0
11	B	1	2	1	1	1	1	1	1	1	1	1	0	0
12	P	0	0	0	0	0	0	0	0	0	0	0	1	0

From:
Needleman &
Wunsch (1970)

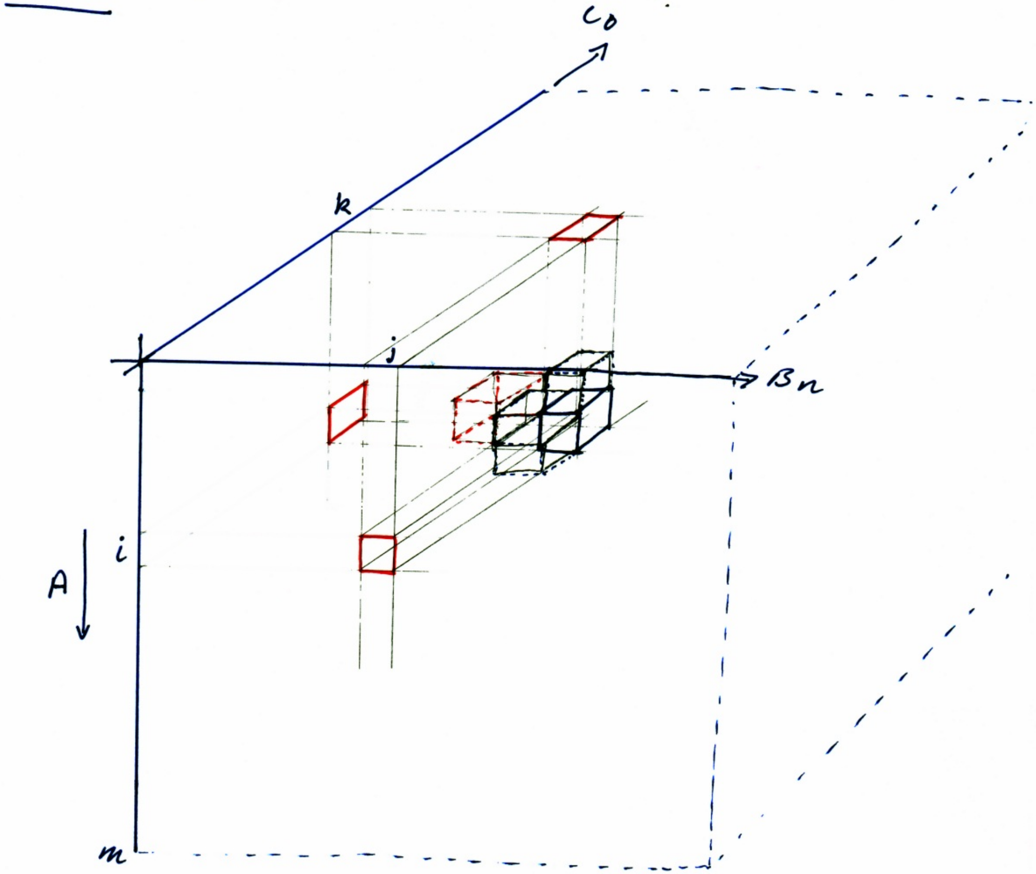
Multiple Sequence Alignment

Extension of two-sequence dynamic programming

For three sequences

Need a 3-dimensional array

3- Sequences?



3- Way Dynamic programming.

- gets complicated! - AND HARD TO DRAW!

BUT NOT IMPOSSIBLE

For *n* sequences?

Need an *n*-dimensional array…

Dynamic programming for >3 sequences

- Need an N-dimensional “hypercube”
- Very complex
- Very memory intensive
- Very CPU intensive

- e.g. to align 100 sequences of length 100.
  Need to store 100^{100} bytes.
  i.e. A BIG NUMBER!

- NOT PRACTICAL

Alternatives to dynamic programming

- Genetic Algorithms
  - Simulate process of “evolution”, but for protein sequence alignments
  - Mutation/recombination of alignments
  - Has been implemented in the SAGA program
- STILL IMPRACTICAL for most use.

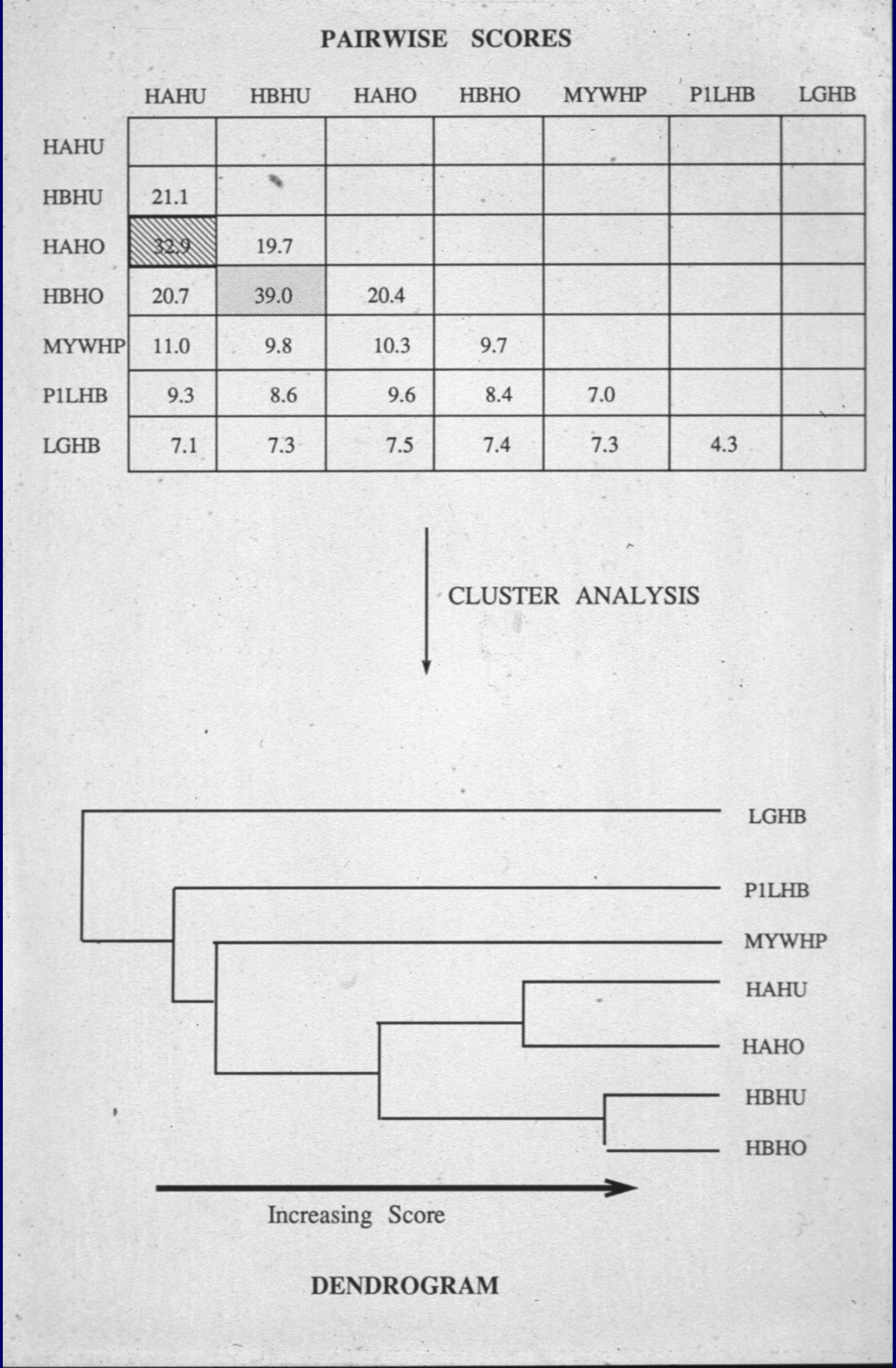
Hierarchical multiple alignment

- Compare all pairs of sequences
- Generate a guide tree or dendrogram
- Follow tree from leaves to root, building the alignment as you go.

- Virtually all current programs use this approach

- Most popular program is CLUSTAL. More recent and often more accurate programs are:
  - probcons, mafft and muscle…

Example: Alignment of 7 sequences with identifier codes HAHU, HBHU etc.



PAIRWISE SCORES

	HAHU	HBHU	HAHO	HBHO	MYWHP	P1LHB	LGHB
HAHU							
HBHU	21.1						
HAHO	32.9	19.7					
HBHO	20.7	39.0	20.4				
MYWHP	11.0	9.8	10.3	9.7			
P1LHB	9.3	8.6	9.6	8.4	7.0		
LGHB	7.1	7.3	7.5	7.4	7.3	4.3	

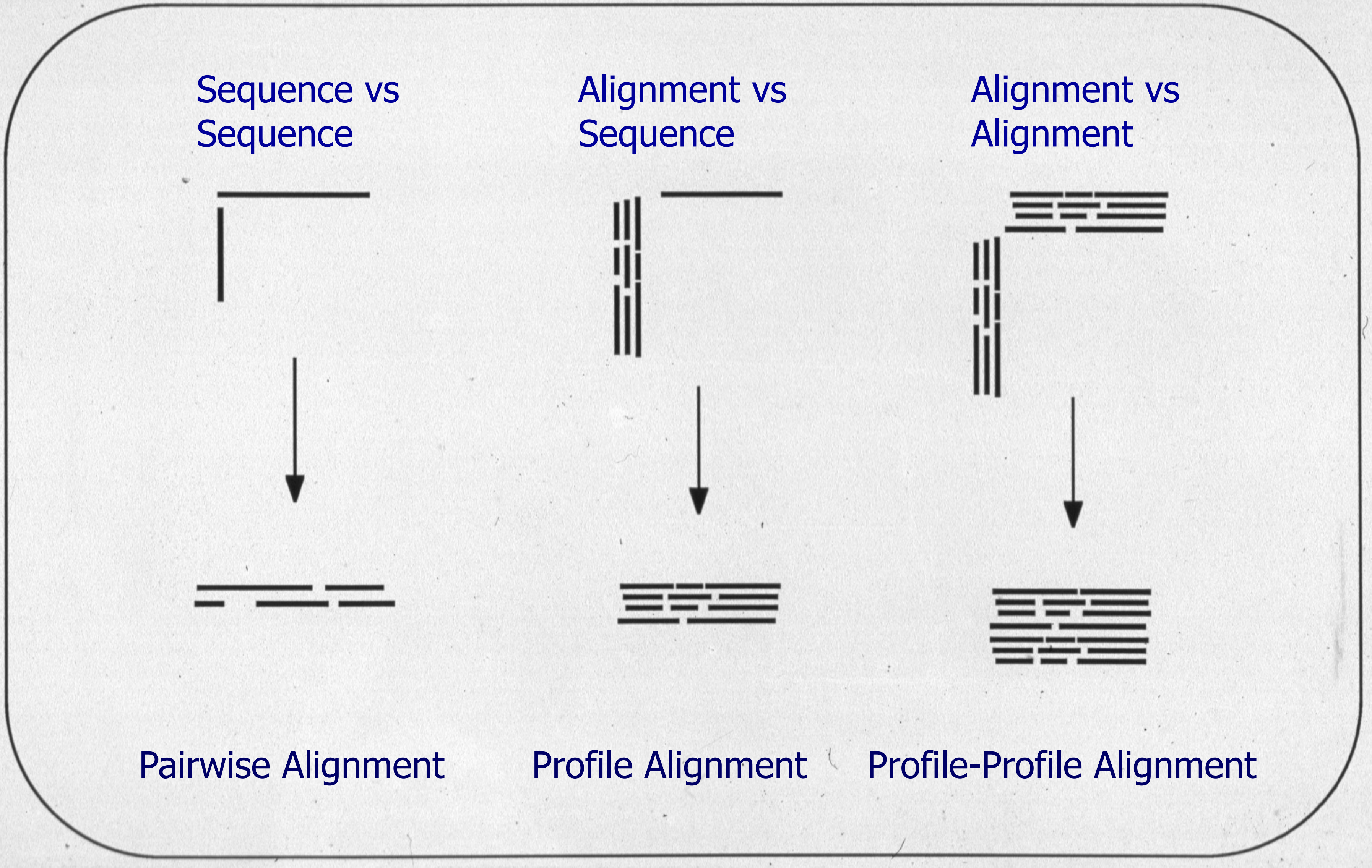
CLUSTER ANALYSIS

DENDROGRAM

“Single linkage” dendrogram.

1. Most similar pair put together
2. Next most similar pair
3. and so on…

When one or both halves of a pair is an existing alignment, then do *profile* comparison.



Sequence vs Sequence

Pairwise Alignment

Alignment vs Sequence

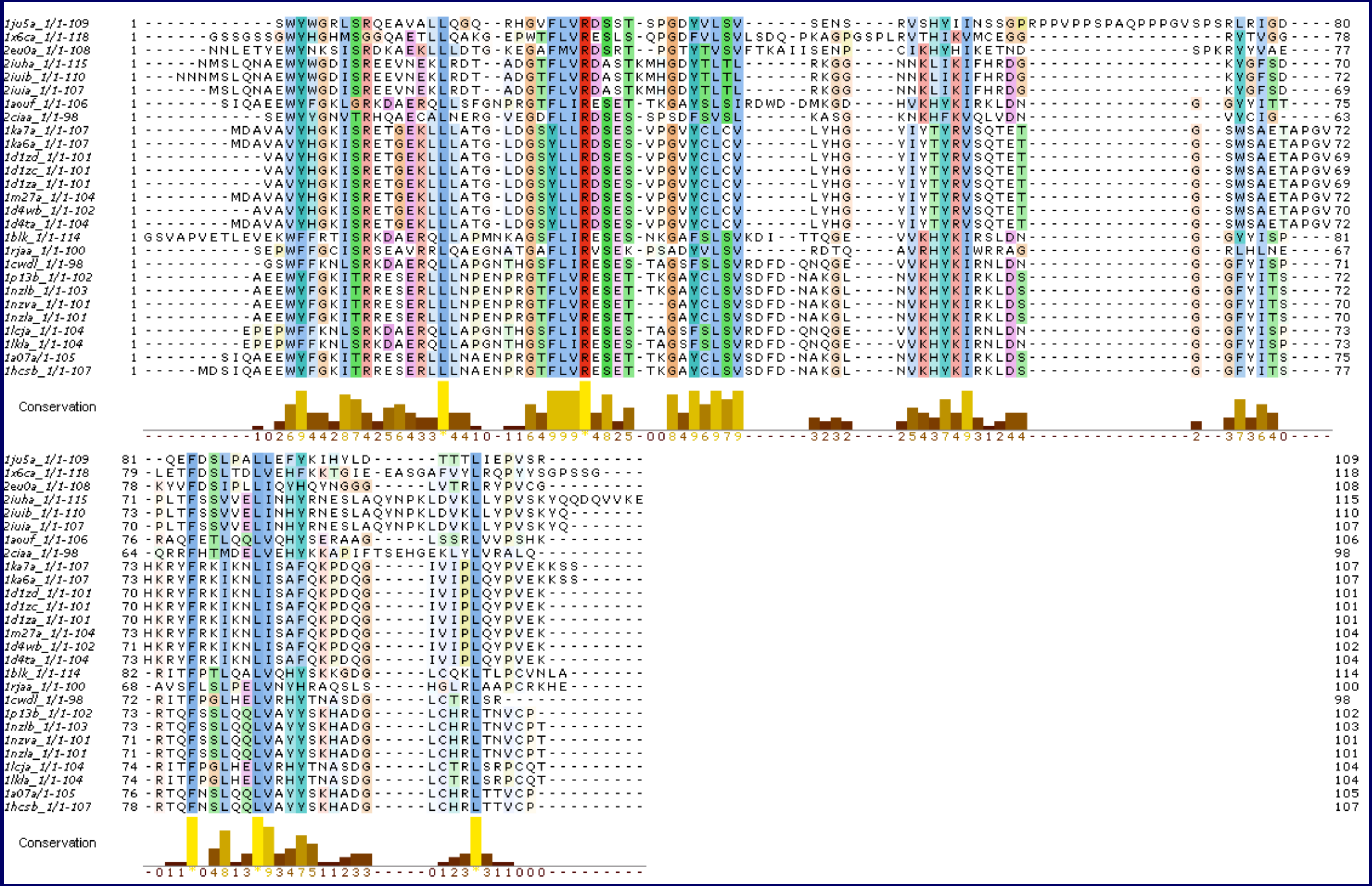
Profile Alignment

Alignment vs Alignment

Profile-Profile Alignment

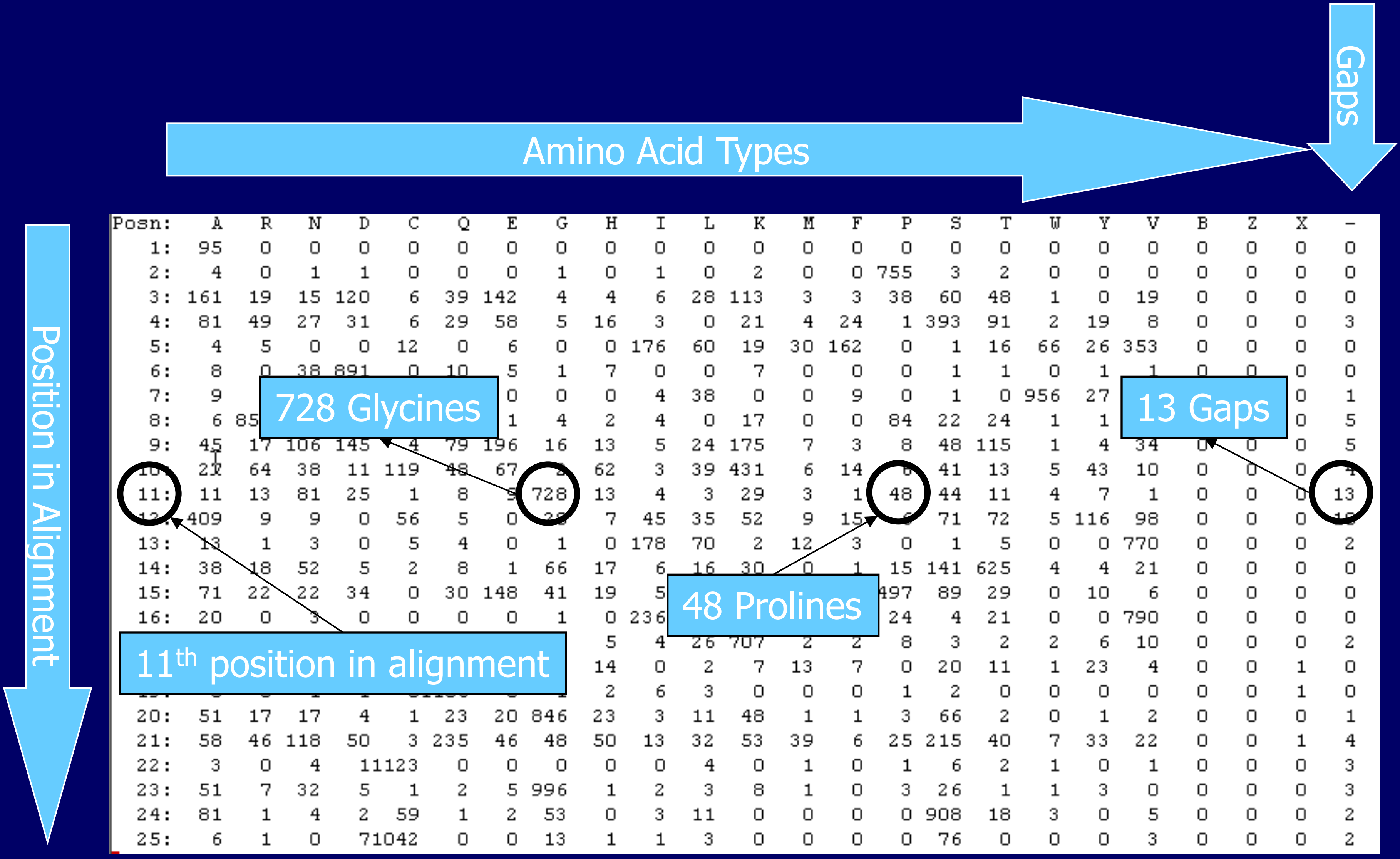
What is a Profile?

Making a profile: Given a multiple sequence alignment…



```
1ju5a_1/1-109    1 ------------SWYWGRLSRQEAVALLQGQ--RHGVFLVRDSST-SPGDYVLSV------SENS----RVSHYIINSSGPRPPVPPSPAQPPPGVSPSRLRIGD----80
1x6ca_1/1-118    1 ------GSSGSSGWYHGHMSGGQAETLLQAKG-EPWTFLVRESLS-QPGDFVLSVLSDQ-PKAGPGSPLRVTHIKVMCEGG------------------RYTVGG----78
2eu0a_1/1-108    1 ------NNLETYEWYNKSISRDKAEKLLLDTG-KEGAFMVRDSRT--PGTYTVSVFTKAIISENP----CIKHYHIKETND---------------SPKRYYVAE----77
2iuha_1/1-115    1 -----NMSLQNAEWYWGDISREEVNEKLRDT--ADGTFLVRDASTKMHGDYTLTL------RKGG----NNKLIKIFHRDG------------------KYGFSD----70
2iuib_1/1-110    1 ---NNNMSLQNAEWYWGDISREEVNEKLRDT--ADGTFLVRDASTKMHGDYTLTL------RKGG----NNKLIKIFHRDG------------------KYGFSD----72
2iuia_1/1-107    1 ------MSLQNAEWYWGDISREEVNEKLRDT--ADGTFLVRDASTKMHGDYTLTL------RKGG----NNKLIKIFHRDG------------------KYGFSD----69
1aouf_1/1-106    1 -------SIQAEEWYFGKLGRKDAERQLLSFGNPRGTFLIRESET-TKGAYSLSIRDWD-DMKGD----HVKHYKIRKLDN---------------G--GYYITT----75
2ciaa_1/1-98     1 -----------SEWYYGNVTRHQAECALNERG-VEGDFLIRDSES-SPSDFSVSL------KASG----KNKHFKVQLVDN------------------VYCIG-----63
1ka7a_1/1-107    1 --------MDAVAVYHGKISRETGEKLLLATG-LDGSYLLRDSES-VPGVYCLCV------LYHG----YIYTYRVSQTET---------------G--SWSAETAPGV72
1ka6a_1/1-107    1 --------MDAVAVYHGKISRETGEKLLLATG-LDGSYLLRDSES-VPGVYCLCV------LYHG----YIYTYRVSQTET---------------G--SWSAETAPGV72
1d1zd_1/1-101    1 -----------VAVYHGKISRETGEKLLLATG-LDGSYLLRDSES-VPGVYCLCV------LYHG----YIYTYRVSQTET---------------G--SWSAETAPGV69
1d1zc_1/1-101    1 -----------VAVYHGKISRETGEKLLLATG-LDGSYLLRDSES-VPGVYCLCV------LYHG----YIYTYRVSQTET---------------G--SWSAETAPGV69
1d1za_1/1-101    1 -----------VAVYHGKISRETGEKLLLATG-LDGSYLLRDSES-VPGVYCLCV------LYHG----YIYTYRVSQTET---------------G--SWSAETAPGV69
1m27a_1/1-104    1 --------MDAVAVYHGKISRETGEKLLLATG-LDGSYLLRDSES-VPGVYCLCV------LYHG----YIYTYRVSQTET---------------G--SWSAETAPGV72
1d4wb_1/1-102    1 ----------AVAVYHGKISRETGEKLLLATG-LDGSYLLRDSES-VPGVYCLCV------LYHG----YIYTYRVSQTET---------------G--SWSAETAPGV70
1d4ta_1/1-104    1 --------MDAVAVYHGKISRETGEKLLLATG-LDGSYLLRDSES-VPGVYCLCV------LYHG----YIYTYRVSQTET---------------G--SWSAETAPGV72
1blk_1/1-114     1 GSVAPVETLEVEKWFFRTISRKDAERQLLAPMNKAGSFLIRESES-NKGAFSLSVKDI--TTQGE----VVKHYKIRSLDN---------------G--GYYISP----81
1rjaa_1/1-100    1 ----------SEPWFFGCISRSEAVRRLQAEGNATGAFLIRVSEK-PSADYVLSV------RDTQ----AVRHYKIWRRAG---------------G--RLHLNE----67
1cwdl_1/1-98     1 -----------GSWFFKNLSRKDAERQLLAPGNTHGSFLIRESES-TAGSFSLSVRDFD-QNQGE----VVKHYKIRNLDN---------------G--GFYISP----71
1p13b_1/1-102    1 ----------AEEWYFGKITRRESERLLLNPENPRGTFLVRESET-TKGAYCLSVSDFD-NAKGL----NVKHYKIRKLDS---------------G--GFYITS----72
1nzlb_1/1-103    1 ----------AEEWYFGKITRRESERLLLNPENPRGTFLVRESET-TKGAYCLSVSDFD-NAKGL----NVKHYKIRKLDS---------------G--GFYITS----72
1nzva_1/1-101    1 ----------AEEWYFGKITRRESERLLLNPENPRGTFLVRESET---GAYCLSVSDFD-NAKGL----NVKHYKIRKLDS---------------G--GFYITS----70
1nzla_1/1-101    1 ----------AEEWYFGKITRRESERLLLNPENPRGTFLVRESET---GAYCLSVSDFD-NAKGL----NVKHYKIRKLDS---------------G--GFYITS----70
1lcja_1/1-104    1 ---------EPEPWFFKNLSRKDAERQLLAPGNTHGSFLIRESES-TAGSFSLSVRDFD-QNQGE----VVKHYKIRNLDN---------------G--GFYISP----73
1lkla_1/1-104    1 ---------EPEPWFFKNLSRKDAERQLLAPGNTHGSFLIRESES-TAGSFSLSVRDFD-QNQGE----VVKHYKIRNLDN---------------G--GFYISP----73
1a07a/1-105      1 -------SIQAEEWYFGKITRRESERLLLNAENPRGTFLVRESET-TKGAYCLSVSDFD-NAKGL----NVKHYKIRKLDS---------------G--GFYITS----75
1hcsb_1/1-107    1 -----MDSIQAEEWYFGKITRRESERLLLNAENPRGTFLVRESET-TKGAYCLSVSDFD-NAKGL----NVKHYKIRKLDS---------------G--GFYITS----77

Conservation
                   ----------10269442874256433*4410-1164999*4825-008496979------3232----254374931244---------------2--373640----

1ju5a_1/1-109   81 --QEFDSLPALLEFYKIHYLD------TTTLIEPVSR---------    109
1x6ca_1/1-118   79 -LETFDSLTDLVEHFKKTGIE-EASGAFVYLRQPYYSGPSSG----    118
2eu0a_1/1-108   78 -KYVFDSIPLLIQYHQYNGGG-----LVTRLRYPVCG---------    108
2iuha_1/1-115   71 -PLTFSSVVELINHYRNESLAQYNPKLDVKLLYPVSKYQQDQVVKE    115
2iuib_1/1-110   73 -PLTFSSVVELINHYRNESLAQYNPKLDVKLLYPVSKYQ-------    110
2iuia_1/1-107   70 -PLTFSSVVELINHYRNESLAQYNPKLDVKLLYPVSKYQ-------    107
1aouf_1/1-106   76 -RAQFETLQQLVQHYSERAAG-----LSSRLVVPSHK---------    106
2ciaa_1/1-98    64 -QRRFHTMDELVEHYKKAPIFTSEHGEKLYLVRALQ----------    98
1ka7a_1/1-107   73 HKRYFRKIKNLISAFQKPDQG-----IVIPLQYPVEKKSS------    107
1ka6a_1/1-107   73 HKRYFRKIKNLISAFQKPDQG-----IVIPLQYPVEKKSS------    107
1d1zd_1/1-101   70 HKRYFRKIKNLISAFQKPDQG-----IVIPLQYPVEK---------    101
1d1zc_1/1-101   70 HKRYFRKIKNLISAFQKPDQG-----IVIPLQYPVEK---------    101
1d1za_1/1-101   70 HKRYFRKIKNLISAFQKPDQG-----IVIPLQYPVEK---------    101
1m27a_1/1-104   73 HKRYFRKIKNLISAFQKPDQG-----IVIPLQYPVEK---------    104
1d4wb_1/1-102   71 HKRYFRKIKNLISAFQKPDQG-----IVIPLQYPVEK---------    102
1d4ta_1/1-104   73 HKRYFRKIKNLISAFQKPDQG-----IVIPLQYPVEK---------    104
1blk_1/1-114    82 -RITFPTLQALVQHYSKKGDG-----LCQKLTLPCVNLA-------    114
1rjaa_1/1-100   68 -AVSFLSLPELVNYHRAQSLS-----HGLRLAAPCRKHE-------    100
1cwdl_1/1-98    72 -RITFPGLHELVRHYTNASDG-----LCTRLSR-------------    98
1p13b_1/1-102   73 -RTQFSSLQQLVAYYSKHADG-----LCHRLTNVCP----------    102
1nzlb_1/1-103   73 -RTQFSSLQQLVAYYSKHADG-----LCHRLTNVCPT---------    103
1nzva_1/1-101   71 -RTQFSSLQQLVAYYSKHADG-----LCHRLTNVCPT---------    101
1nzla_1/1-101   71 -RTQFSSLQQLVAYYSKHADG-----LCHRLTNVCPT---------    101
1lcja_1/1-104   74 -RITFPGLHELVRHYTNASDG-----LCTRLSRPCQT---------    104
1lkla_1/1-104   74 -RITFPGLHELVRHYTNASDG-----LCTRLSRPCQT---------    104
1a07a/1-105     76 -RTQFNSLQQLVAYYSKHADG-----LCHRLTTVCP----------    105
1hcsb_1/1-107   78 -RTQFNSLQQLVAYYSKHADG-----LCHRLTTVCP----------    107

Conservation
                   -011*04813*93475112 33-----0123*311000---------
```

Multiple Alignment Frequency Profile



Amino Acid Types

Gaps

Position in Alignment

728 Glycines

13 Gaps

48 Prolines

11th position in alignment

Posn:	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	-
1:	95	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2:	4	0	1	1	0	0	0	1	0	1	0	2	0	0	755	3	2	0	0	0	0	0	0	0
3:	161	19	15	120	6	39	142	4	4	6	28	113	3	3	38	60	48	1	0	19	0	0	0	0
4:	81	49	27	31	6	29	58	5	16	3	0	21	4	24	1	393	91	2	19	8	0	0	0	3
5:	4	5	0	0	12	0	6	0	0	176	60	19	30	162	0	1	16	66	26	353	0	0	0	0
6:	8	0	38	891	0	10	5	1	7	0	0	7	0	0	0	1	1	0	1	1	0	0	0	0
7:	9	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	0	0	4	38	0	0	9	0	1	0	956	27	[illegible]	[illegible]	[illegible]	0	1
8:	6	85	[illegible]	[illegible]	[illegible]	[illegible]	1	4	2	4	0	17	0	0	84	22	24	1	1	[illegible]	[illegible]	[illegible]	0	5
9:	45	17	106	145	4	79	196	16	13	5	24	175	7	3	8	48	115	1	4	34	0	0	0	5
[illegible]	[illegible]	64	38	11	119	48	67	2	62	3	39	431	6	14	6	41	13	5	43	10	0	0	0	4
11:	11	13	81	25	1	8	9	728	13	4	3	29	3	1	48	44	11	4	7	1	0	0	0	13
12:	409	9	9	0	56	5	0	[illegible]	7	45	35	52	9	15	6	71	72	5	116	98	0	0	0	[illegible]
13:	13	1	3	0	5	4	0	1	0	178	70	2	12	3	0	1	5	0	0	770	0	0	0	2
14:	38	18	52	5	2	8	1	66	17	6	16	30	0	1	15	141	625	4	4	21	0	0	0	0
15:	71	22	22	34	0	30	148	41	19	5	[illegible]	[illegible]	[illegible]	[illegible]	497	89	29	0	10	6	0	0	0	0
16:	20	0	3	0	0	0	0	1	0	236	[illegible]	[illegible]	[illegible]	[illegible]	24	4	21	0	0	790	0	0	0	0
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	5	4	26	707	2	2	8	3	2	2	6	10	0	0	0	2
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	14	0	2	7	13	7	0	20	11	1	23	4	0	0	1	0
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2	6	3	0	0	0	1	2	0	0	0	0	0	0	1	0
20:	51	17	17	4	1	23	20	846	23	3	11	48	1	1	3	66	2	0	1	2	0	0	0	1
21:	58	46	118	50	3	235	46	48	50	13	32	53	39	6	25	215	40	7	33	22	0	0	1	4
22:	3	0	4	1	1123	0	0	0	0	0	4	0	1	0	1	6	2	1	0	1	0	0	0	3
23:	51	7	32	5	1	2	5	996	1	2	3	8	1	0	3	26	1	1	3	0	0	0	0	3
24:	81	1	4	2	59	1	2	53	0	3	11	0	0	0	0	908	18	3	0	5	0	0	0	2
25:	6	1	0	7	1042	0	0	13	1	1	3	0	0	0	0	76	0	0	0	3	0	0	0	2

Convert frequency profile into log-odds profile

In words:

log (proportion of a particular amino acid type at a position
divided by
proportion of that amino acid in the whole alignment)

Results in a
Negative number when amino acid is less common at a position than in the alignment as a whole.
or a
Positive number if the amino acid is more common at a position than in the alignment as a whole.

Conversion is usually more complex than this because you have to deal with the absence of amino acids at a position. This is done by taking background scores from a pair-score matrix like BLOSUM.

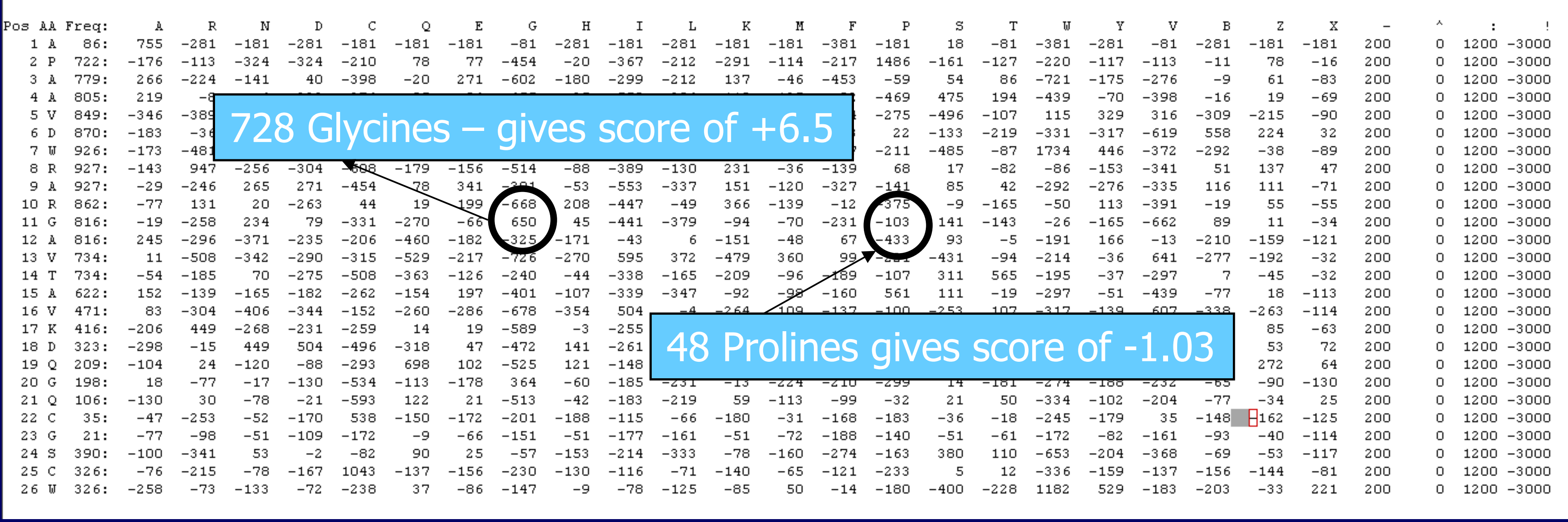
Example

- Alignment of 30 sequences each of 100 amino acids to give a total of 3000 amino acids
- Position 97 of the alignment has 20 prolines
- There are 300 prolines in the alignment as a whole
- (20/30)/(300/3000) = 20/3 = 6.67
- log(6.67) = 0.82
- So, score in profile for proline at position 97 is 0.82

- This is sometimes called a log-likelihood ratio

Log score profile

- In this example log values are multiplied by 100 to allow for integer arithmetic which is faster on most computers.



Pos	AA	Freq:	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	-	^	:	!
1	A	86:	755	-281	-181	-281	-181	-181	-181	-81	-281	-181	-281	-181	-181	-381	-181	18	-81	-381	-281	-81	-281	-181	-181	200	0	1200	-3000
2	P	722:	-176	-113	-324	-324	-210	78	77	-454	-20	-367	-212	-291	-114	-217	1486	-161	-127	-220	-117	-113	-11	78	-16	200	0	1200	-3000
3	A	779:	266	-224	-141	40	-398	-20	271	-602	-180	-299	-212	137	-46	-453	-59	54	86	-721	-175	-276	-9	61	-83	200	0	1200	-3000
4	A	805:	219	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-469	475	194	-439	-70	-398	-16	19	-69	200	0	1200	-3000
5	V	849:	-346	-389	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-275	-496	-107	115	329	316	-309	-215	-90	200	0	1200	-3000
6	D	870:	-183	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	22	-133	-219	-331	-317	-619	558	224	32	200	0	1200	-3000
7	W	926:	-173	-481	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-211	-485	-87	1734	446	-372	-292	-38	-89	200	0	1200	-3000
8	R	927:	-143	947	-256	-304	-608	-179	-156	-514	-88	-389	-130	231	-36	-139	68	17	-82	-86	-153	-341	51	137	47	200	0	1200	-3000
9	A	927:	-29	-246	265	271	-454	78	341	-381	-53	-553	-337	151	-120	-327	-141	85	42	-292	-276	-335	116	111	-71	200	0	1200	-3000
10	R	862:	-77	131	20	-263	44	19	199	-668	208	-447	-49	366	-139	-12	-375	-9	-165	-50	113	-391	-19	55	-55	200	0	1200	-3000
11	G	816:	-19	-258	234	79	-331	-270	-66	650	45	-441	-379	-94	-70	-231	-103	141	-143	-26	-165	-662	89	11	-34	200	0	1200	-3000
12	A	816:	245	-296	-371	-235	-206	-460	-182	-325	-171	-43	6	-151	-48	67	-433	93	-5	-191	166	-13	-210	-159	-121	200	0	1200	-3000
13	V	734:	11	-508	-342	-290	-315	-529	-217	-726	-270	595	372	-479	360	99	-281	-431	-94	-214	-36	641	-277	-192	-32	200	0	1200	-3000
14	T	734:	-54	-185	70	-275	-508	-363	-126	-240	-44	-338	-165	-209	-96	-189	-107	311	565	-195	-37	-297	7	-45	-32	200	0	1200	-3000
15	A	622:	152	-139	-165	-182	-262	-154	197	-401	-107	-339	-347	-92	-98	-160	561	111	-19	-297	-51	-439	-77	18	-113	200	0	1200	-3000
16	V	471:	83	-304	-406	-344	-152	-260	-286	-678	-354	504	-4	-264	109	-137	-100	-253	107	-317	-139	607	-338	-263	-114	200	0	1200	-3000
17	K	416:	-206	449	-268	-231	-259	14	19	-589	-3	-255	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	85	-63	200	0	1200	-3000
18	D	323:	-298	-15	449	504	-496	-318	47	-472	141	-261	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	53	72	200	0	1200	-3000
19	Q	209:	-104	24	-120	-88	-293	698	102	-525	121	-148	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	272	64	200	0	1200	-3000
20	G	198:	18	-77	-17	-130	-534	-113	-178	364	-60	-185	-231	-13	-224	-210	-299	14	-181	-274	-188	-232	-65	-90	-130	200	0	1200	-3000
21	Q	106:	-130	30	-78	-21	-593	122	21	-513	-42	-183	-219	59	-113	-99	-32	21	50	-334	-102	-204	-77	-34	25	200	0	1200	-3000
22	C	35:	-47	-253	-52	-170	538	-150	-172	-201	-188	-115	-66	-180	-31	-168	-183	-36	-18	-245	-179	35	-148	-162	-125	200	0	1200	-3000
23	G	21:	-77	-98	-51	-109	-172	-9	-66	-151	-51	-177	-161	-51	-72	-188	-140	-51	-61	-172	-82	-161	-93	-40	-114	200	0	1200	-3000
24	S	390:	-100	-341	53	-2	-82	90	25	-57	-153	-214	-333	-78	-160	-274	-163	380	110	-653	-204	-368	-69	-53	-117	200	0	1200	-3000
25	C	326:	-76	-215	-78	-167	1043	-137	-156	-230	-130	-116	-71	-140	-65	-121	-233	5	12	-336	-159	-137	-156	-144	-81	200	0	1200	-3000
26	W	326:	-258	-73	-133	-72	-238	37	-86	-147	-9	-78	-125	-85	50	-14	-180	-400	-228	1182	529	-183	-203	-33	221	200	0	1200	-3000

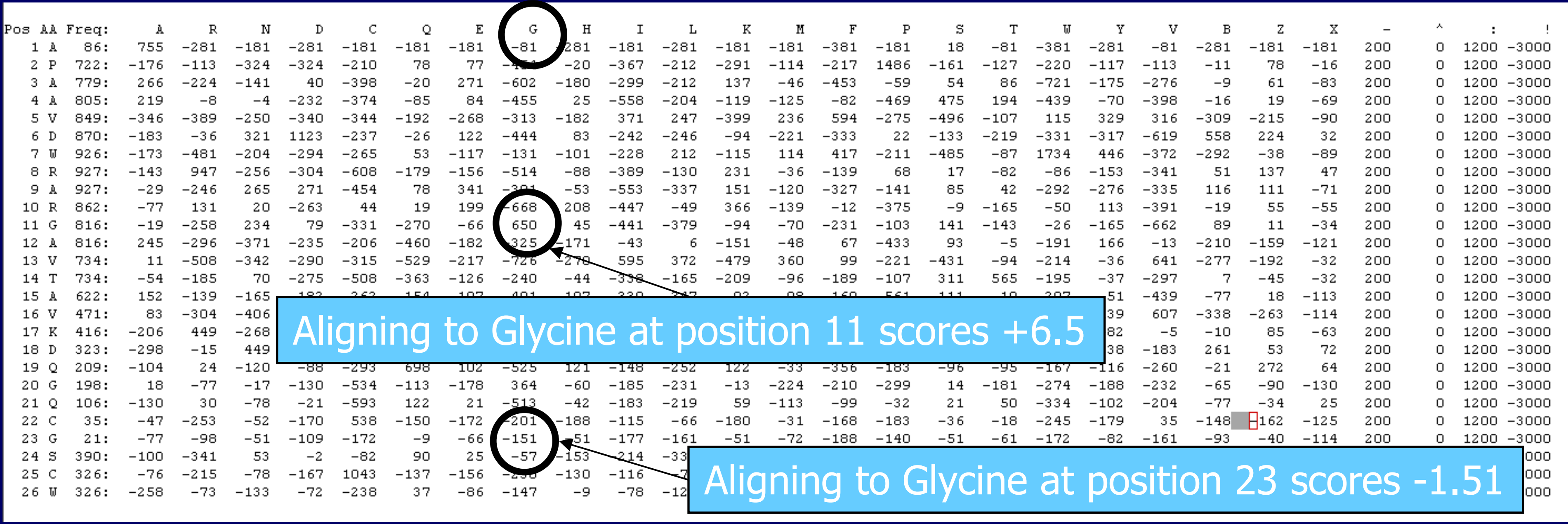
728 Glycines – gives score of +6.5

48 Prolines gives score of -1.03

How is a profile used in alignment?

- Rather than getting the score for aligning a particular residue at a position from the BLOSUM matrix take it from the profile.

Profiles give position-specific scoring



```
Pos AA Freq:      A      R      N      D      C      Q      E      G      H      I      L      K      M      F      P      S      T      W      Y      V      B      Z      X      -      ^      :      !
  1 A    86:    755   -281   -181   -281   -181   -181   -181    -81   -281   -181   -281   -181   -181   -381   -181     18    -81   -381   -281    -81   -281   -181   -181    200      0   1200  -3000
  2 P   722:   -176   -113   -324   -324   -210     78     77   -454    -20   -367   -212   -291   -114   -217   1486   -161   -127   -220   -117   -113    -11     78    -16    200      0   1200  -3000
  3 A   779:    266   -224   -141     40   -398    -20    271   -602   -180   -299   -212    137    -46   -453    -59     54     86   -721   -175   -276     -9     61    -83    200      0   1200  -3000
  4 A   805:    219     -8     -4   -232   -374    -85     84   -455     25   -558   -204   -119   -125    -82   -469    475    194   -439    -70   -398    -16     19    -69    200      0   1200  -3000
  5 V   849:   -346   -389   -250   -340   -344   -192   -268   -313   -182    371    247   -399    236    594   -275   -496   -107    115    329    316   -309   -215    -90    200      0   1200  -3000
  6 D   870:   -183    -36    321   1123   -237    -26    122   -444     83   -242   -246    -94   -221   -333     22   -133   -219   -331   -317   -619    558    224     32    200      0   1200  -3000
  7 W   926:   -173   -481   -204   -294   -265     53   -117   -131   -101   -228    212   -115    114    417   -211   -485    -87   1734    446   -372   -292    -38    -89    200      0   1200  -3000
  8 R   927:   -143    947   -256   -304   -608   -179   -156   -514    -88   -389   -130    231    -36   -139     68     17    -82    -86   -153   -341     51    137     47    200      0   1200  -3000
  9 A   927:    -29   -246    265    271   -454     78    341   -391    -53   -553   -337    151   -120   -327   -141     85     42   -292   -276   -335    116    111    -71    200      0   1200  -3000
 10 R   862:    -77    131     20   -263     44     19    199   -668    208   -447    -49    366   -139    -12   -375     -9   -165    -50    113   -391    -19     55    -55    200      0   1200  -3000
 11 G   816:    -19   -258    234     79   -331   -270    -66    650     45   -441   -379    -94    -70   -231   -103    141   -143    -26   -165   -662     89     11    -34    200      0   1200  -3000
 12 A   816:    245   -296   -371   -235   -206   -460   -182   -325   -171    -43      6   -151    -48     67   -433     93     -5   -191    166    -13   -210   -159   -121    200      0   1200  -3000
 13 V   734:     11   -508   -342   -290   -315   -529   -217   -726   -270    595    372   -479    360     99   -221   -431    -94   -214    -36    641   -277   -192    -32    200      0   1200  -3000
 14 T   734:    -54   -185     70   -275   -508   -363   -126   -240    -44   -338   -165   -209    -96   -189   -107    311    565   -195    -37   -297      7    -45    -32    200      0   1200  -3000
 15 A   622:    152   -139   -165                                                                                                     51   -439    -77     18   -113    200      0   1200  -3000
 16 V   471:     83   -304   -406                                                                                                     39    607   -338   -263   -114    200      0   1200  -3000
 17 K   416:   -206    449   -268                                                                                                     82     -5    -10     85    -63    200      0   1200  -3000
 18 D   323:   -298    -15    449                                                                                                     38   -183    261     53     72    200      0   1200  -3000
 19 Q   209:   -104     24   -120    -88   -293    698    102   -525    121   -148   -252    122    -33   -356   -183    -96    -95   -167   -116   -260    -21    272     64    200      0   1200  -3000
 20 G   198:     18    -77    -17   -130   -534   -113   -178    364    -60   -185   -231    -13   -224   -210   -299     14   -181   -274   -188   -232    -65    -90   -130    200      0   1200  -3000
 21 Q   106:   -130     30    -78    -21   -593    122     21   -513    -42   -183   -219     59   -113    -99    -32     21     50   -334   -102   -204    -77    -34     25    200      0   1200  -3000
 22 C    35:    -47   -253    -52   -170    538   -150   -172   -201   -188   -115    -66   -180    -31   -168   -183    -36    -18   -245   -179     35   -148   -162   -125    200      0   1200  -3000
 23 G    21:    -77    -98    -51   -109   -172     -9    -66   -151    -51   -177   -161    -51    -72   -188   -140    -51    -61   -172    -82   -161    -93    -40   -114    200      0   1200  -3000
 24 S   390:   -100   -341     53     -2    -82     90     25    -57   -153   -214    -33                                                                                                           000
 25 C   326:    -76   -215    -78   -167   1043   -137   -156   -290   -130   -116     -7                                                                                                           000
 26 W   326:   -258    -73   -133    -72   -238     37    -86   -147     -9    -78    -12                                                                                                           000
```

Aligning to Glycine at position 11 scores +6.5

Aligning to Glycine at position 23 scores -1.51

This emphasises position-specific features of the protein family

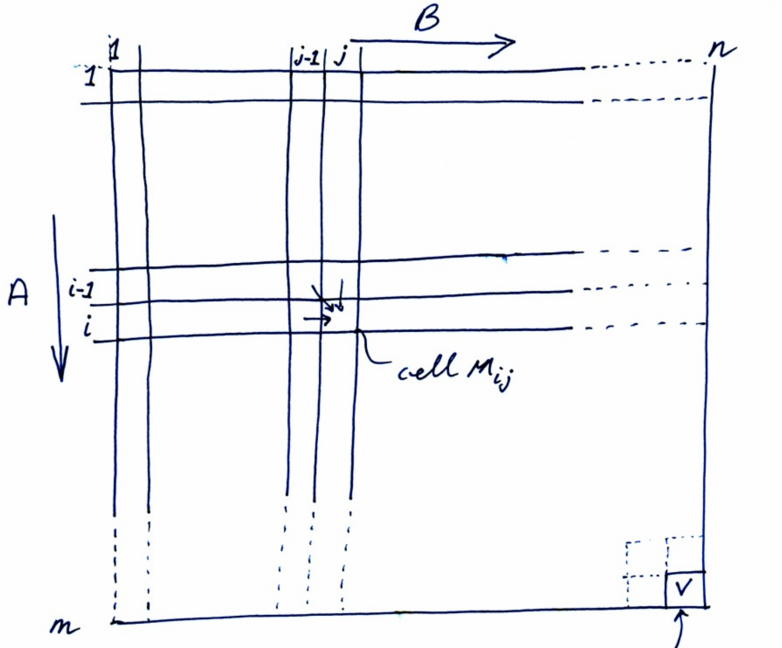
Compared to Gly-Gly score of 0.6 in the BLOSUM62 matrix.

When either A or B is a profile, the score comes from the profile rather than the BLOSUM matrix.

If A and B are both profiles, then the score is obtained by combining the scores from the two profiles (Exactly how is beyond this lecture)

DYNAMIC PROGRAMMING

Matrix M



Sequences $A_{1....m}$ $B_{1.....n}$ $M_{1...m,1...n}$

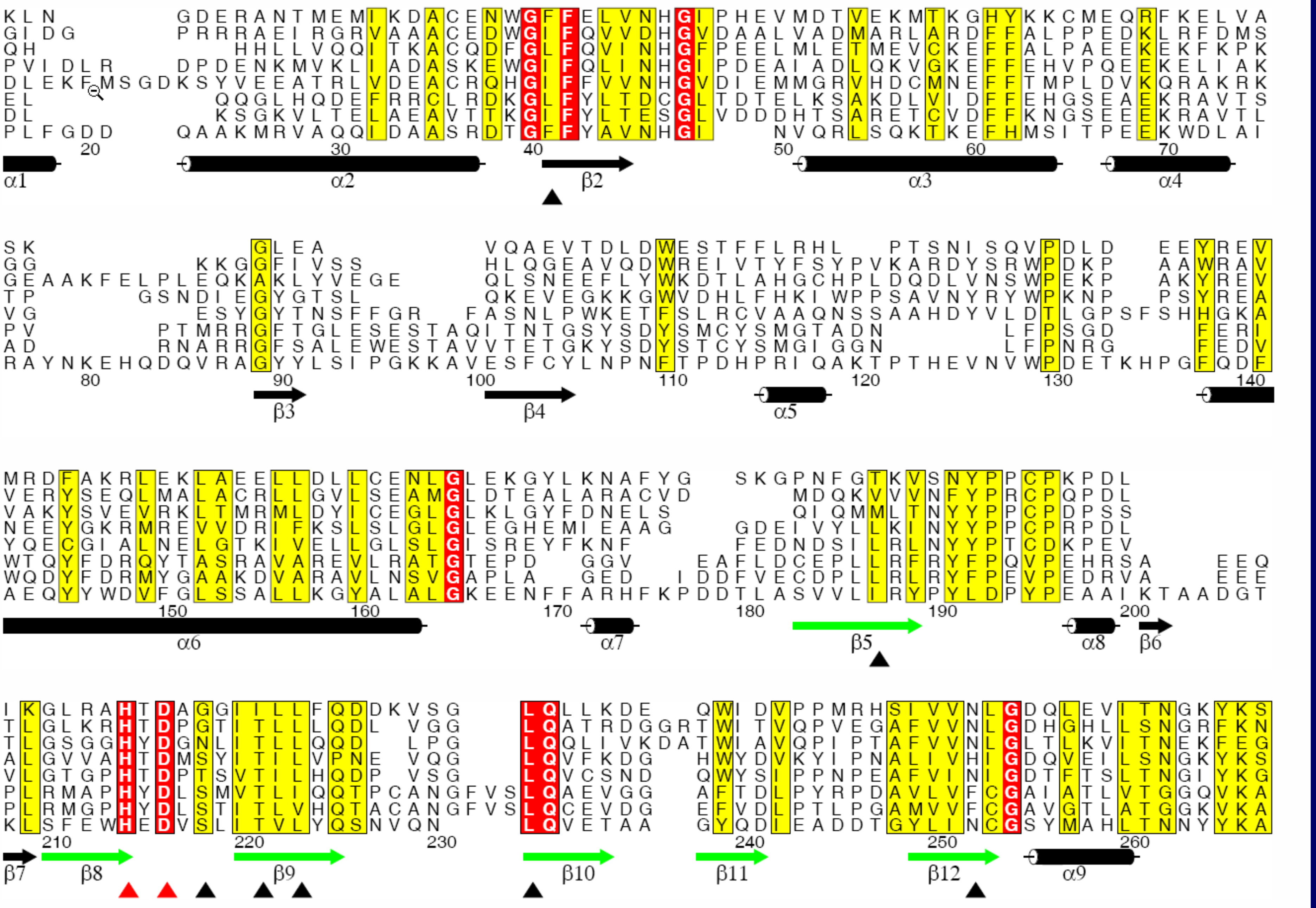
V = total score for alignment of A&B

$$\text{Value of } M_{ij} = \max \begin{cases} M_{i,j-1} - pen \\ M_{j,i-1} - pen \\ M_{i,j} + Score_{A_i,B_j} \end{cases}$$

where pen = gap penalty and

$Score_{A_i B_j}$ comes from BLOSUM or PAM Substitution matrix (for proteins)

For 2 sequences mn steps.



```
KLN      GDERANTMEMIKDACENWGFFELVNHGIPHEVMDTVEKMTKGHYKKCMEQRFKELVA
GIDG     PRRRAEIRGRVAAACEDWGIFQVVDHGVDAALVADMARLARDFFALPPEDKLRFDMS
QH          HHLLVQQITKACQDFGLFQVINHGFPEELMLETMEVCKEFFALPAEEKEKFKPK
PVIDLR   DPDENKMVKLIADASKEWGIFQLINHGIPDEAIADLQKVGKEFFEHVPQEEKELIAK
DLEKFMSGDKSYVEEATRLVDEACRQHGIFFVVNHGVDIEMMGRVHDCMNEFFTMPLDVKQRAKRK
EL         QQGLHQDEFRRCLRDKGLFYLTDCGLTDTELKSAKDLVIDFFEHGSEAEKRAVTS
DL         KSGKVLTELAEAVTTKGIFYLTESGLVDDDHTSARETCVDFFKNGSEEEKRAVTL
PLFGDD   QAAKMRVAQQIDAASRDTGFFYAVNHGI   NVQRLSQKTKEFHMSITPEEKWDLAI
    20           30        40           50        60        70

SK           GLEA        VQAEVTDLDWESTFFLRHL  PTSNISQVPDLD  EEYREV
GG        KKGGFIVSS      HLQGEAVQDWREIVTYFSYPVKARDYSRWPDKP  AAWRAV
GEAAKFELPLEQKAKLYVEGE    QLSNEEFLYWKDTLAHGCHPLDQDLVNSWPEKP  AKYREV
TP     GSNDIEGYGTSL      QKEVEGKKGWVDHLFHKIWPPSAVNYRYWPKNP  PSYREA
VG        ESYGYTNSFFGR  FASNLPWKETFSLRCVAAQNSSAAHDYVLDTLGPSFSHHGKA
PV      PTMRRGFTGLESESTAQITNTGSYSDYSMCYSMGTADN      LFPSGD    FERI
AD      RNARRGFSALEWESTAVVTETGKYSDYSTCYSMGIGGN      LFPNRG    FEDV
RAYNKEHQDQVRAGYYLSIPGKKAVESFCYLNPNFTPDHPRIQAKTPTHEVNVWPDETKHPGFQDF
    80        90        100       110       120       130       140

MRDFAKRLEKLAEELLDLLCENLGLEKGYLKNAFYG  SKGPNFGTKVSNYPPCPKPDL
VERYSEQLMALACRLLGVLSEAMGLDTEALARACVD     MDQKVVVNFYPRCPQPDL
VAKYSVEVRKLTMRMLDYICEGLGLKLGYFDNELS      QIQMMLTNYYPPCPDPSS
NEEYGKRMREVVDRIFKSLSLGLGLEGHEMIEAAG   GDEIVYLLKINYYPPCPRPDL
YQECGIALNELGTKIVELLGLSLGISREYFKNF     FEDNDSILRLNYYPTCDKPEV
WTQYFDRQYTASRAVAREVLRATGTEPD  GGV   EAFLDCEPLLRFRYFPQVPEHRSA   EEQ
WQDYFDRMYGAAKDVARAVLNSVGAPLA  GED  IDDFVECDPLLRLRYFPEVPEDRVA   EEE
AEQYYWDVFGLSSALLKGYALALGKEENFFARHFKPDDTLASVVLIRYPYLDPYPEAAIKTAADGT
        150       160       170       180       190       200

IKGLRAHTDAGGIILLFQDDKVSG   LQLLKDE  QWIDVPPMRHSIVVNLGDQLEVITNGKYKS
TLGLKRHTDPGTITLLLQDL VGG   LQATRDGGRTWITVQPVEGAFVVNLGDHGHLLSNGRFKN
TLGSGGHYDGNLITLLQQD  LPG   LQQLIVKDATWIAVQPIPTAFVVNLGLTLKVITNEKFEG
ALGVVAHTDMSYITILVPNE VQG   LQVFKDG  HWYDVKYIPNALIVHIGDQVEILSNGKYKS
VLGTGPHTDPTSVTILHQDP VSG   LQVCSND  QWYSIPPNPEAFVINIGDTFTSLTNGIYKG
PLRMAPHYDLSMVTLIQQTPCANGFVSLQAEVGG  AFTDLPYRPDAVLVFCGAITLVTGGQVKA
PLRMGPHYDLSTITLVHQTACANGFVSLQCEVDG  EFVDLPTLPGAMVVFCGAVGTLATGGKVKA
KLSFEWHEDVSLITVLYQSNVQN    LQVETAA  GYQDIEADDTGYLINCGSYMAHLTNNYYKA
  210       220       230             240       250       260
```

ALSCRIPT output with colouring and secondary structure logos

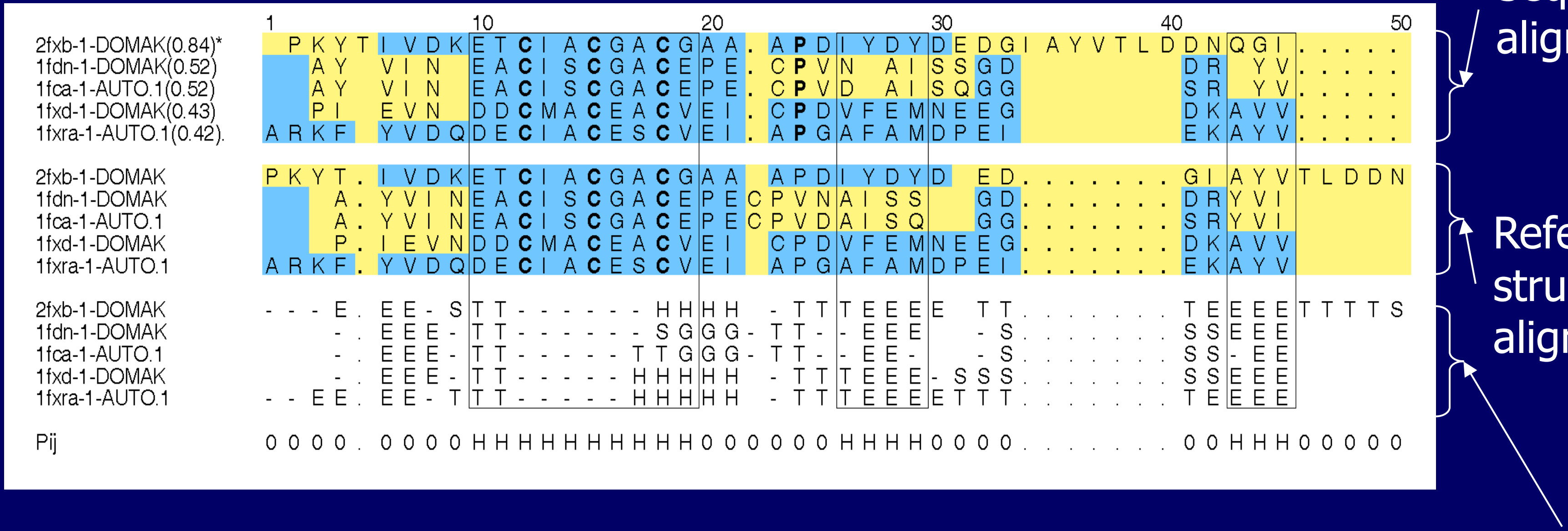
(colour inkjet printer…) Roach *et al*, (1995)

How good are alignments?

Use of reference alignments to see how well sequence alignments work

- OXBench – library of 672 multiple structure alignments
- Software to test how well different methods work
  - Raghava GP, Searle SM, Audley PC, Barber JD, Barton GJ. OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. BMC Bioinformatics. 2003 Oct 10;4:47.

Comparison of Structural sequence alignment to sequence alignment



```
                        1                 10                  20                  30                  40                  50
2fxb-1-DOMAK(0.84)*       P K Y T I V D K E T C I A C G A C G A A . A P D I Y D Y D E D G I A Y V T L D D N Q G I . . . . .
1fdn-1-DOMAK(0.52)          A Y   V I N   E A C I S C G A C E P E . C P V N   A I S S G D               D R   Y V . . . . .
1fca-1-AUTO.1(0.52)         A Y   V I N   E A C I S C G A C E P E . C P V D   A I S Q G G               S R   Y V . . . . .
1fxd-1-DOMAK(0.43)          P I   E V N   D D C M A C E A C V E I . C P D V F E M N E E G               D K A V V . . . . .
1fxra-1-AUTO.1(0.42).   A R K F   Y V D Q D E C I A C E S C V E I . A P G A F A M D P E I               E K A Y V . . . . .

2fxb-1-DOMAK            P K Y T . I V D K E T C I A C G A C G A A   A P D I Y D Y D   E D . . . . . . . G I A Y V T L D D N
1fdn-1-DOMAK                  A . Y V I N E A C I S C G A C E P E C P V N A I S S     G D . . . . . . . D R Y V I
1fca-1-AUTO.1                 A . Y V I N E A C I S C G A C E P E C P V D A I S Q     G G . . . . . . . S R Y V I
1fxd-1-DOMAK                  P . I E V N D D C M A C E A C V E I   C P D V F E M N E E G . . . . . . . D K A V V
1fxra-1-AUTO.1          A R K F . Y V D Q D E C I A C E S C V E I   A P G A F A M D P E I . . . . . . . E K A Y V

2fxb-1-DOMAK            - - - E . E E - S T T - - - - - - H H H H   - T T T E E E E   T T . . . . . . . T E E E E T T T T S
1fdn-1-DOMAK                  - . E E E - T T - - - - - - S G G G - T T - - E E E     - S . . . . . . . S S E E E
1fca-1-AUTO.1                 - . E E E - T T - - - - - T T G G G - T T - - E E -     - S . . . . . . . S S - E E
1fxd-1-DOMAK                  - . E E E - T T - - - - - H H H H H   - T T T E E E - S S S . . . . . . . S S E E E
1fxra-1-AUTO.1          - - E E . E E - T T T - - - - - H H H H H   - T T T E E E E T T T . . . . . . . T E E E E

Pij                     0 0 0 0 . 0 0 0 0 H H H H H H H H H H 0 0 0 0 0 0 H H H H 0 0 0 0 . . . . . . . 0 0 H H H 0 0 0 0 0
```

Sequence alignment

Reference structural alignment

Secondary structure for reference alignment

Boxed regions: STAMP reliably structurally aligned

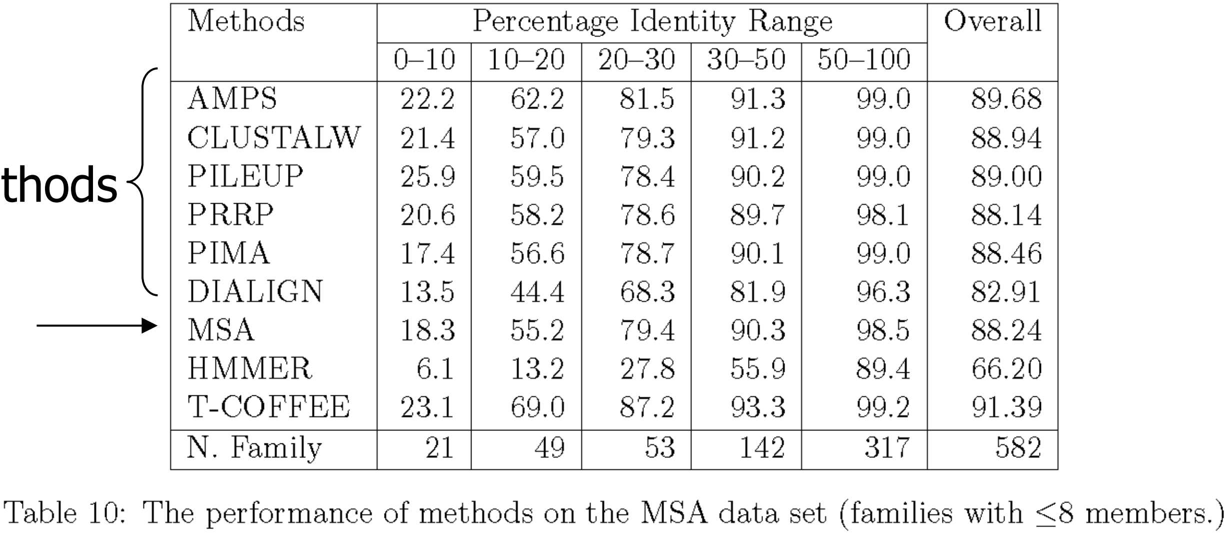
BLUE highlighting: parts of the alignments that are the same.
YELLOW highlighting: parts of the alignment that is different.

Result of comparisons on alignments of 8 sequences or less

Grouped by percentage sequence identity (more on that later)

Hierarchical Methods

N-way DP with corner cutting



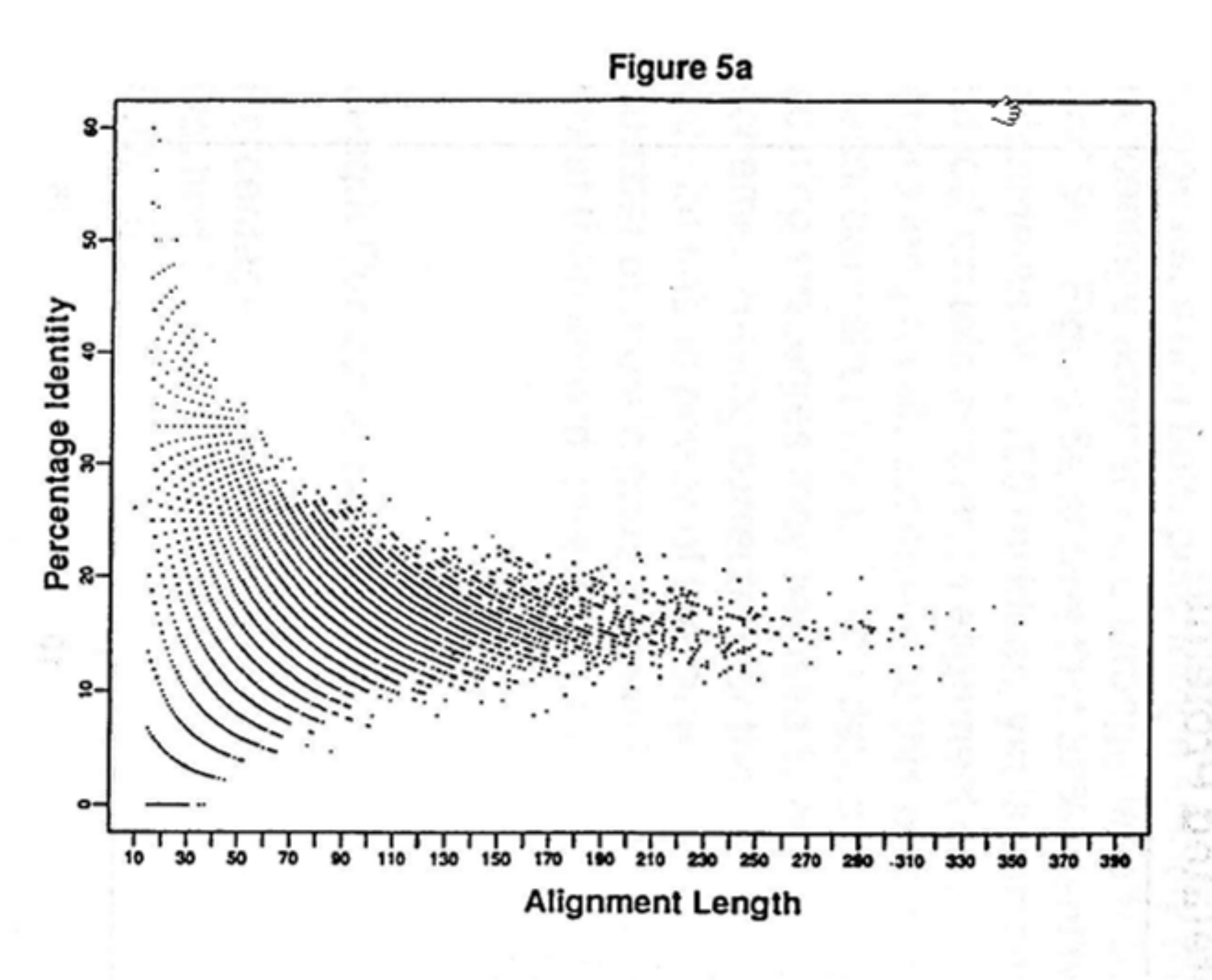
Methods	Percentage Identity Range					Overall
	0–10	10–20	20–30	30–50	50–100	
AMPS	22.2	62.2	81.5	91.3	99.0	89.68
CLUSTALW	21.4	57.0	79.3	91.2	99.0	88.94
PILEUP	25.9	59.5	78.4	90.2	99.0	89.00
PRRP	20.6	58.2	78.6	89.7	98.1	88.14
PIMA	17.4	56.6	78.7	90.1	99.0	88.46
DIALIGN	13.5	44.4	68.3	81.9	96.3	82.91
MSA	18.3	55.2	79.4	90.3	98.5	88.24
HMMER	6.1	13.2	27.8	55.9	89.4	66.20
T-COFFEE	23.1	69.0	87.2	93.3	99.2	91.39
N. Family	21	49	53	142	317	582

Table 10: The performance of methods on the MSA data set (families with ≤8 members.)

How similar do sequences need to be before we can align them reliably?

Percentage identity is strongly length dependent

Pair-wise sequence alignments of proteins known to be unrelated.



Barton, GJ, Proceedings of the CCP4 Study Weekend on Molecular Replacement (31 Jan-1 Feb, 1992)
There is a more recent ref with similar figure in it by Burkhard Rost, but I must find it!

Problems with percentage identity

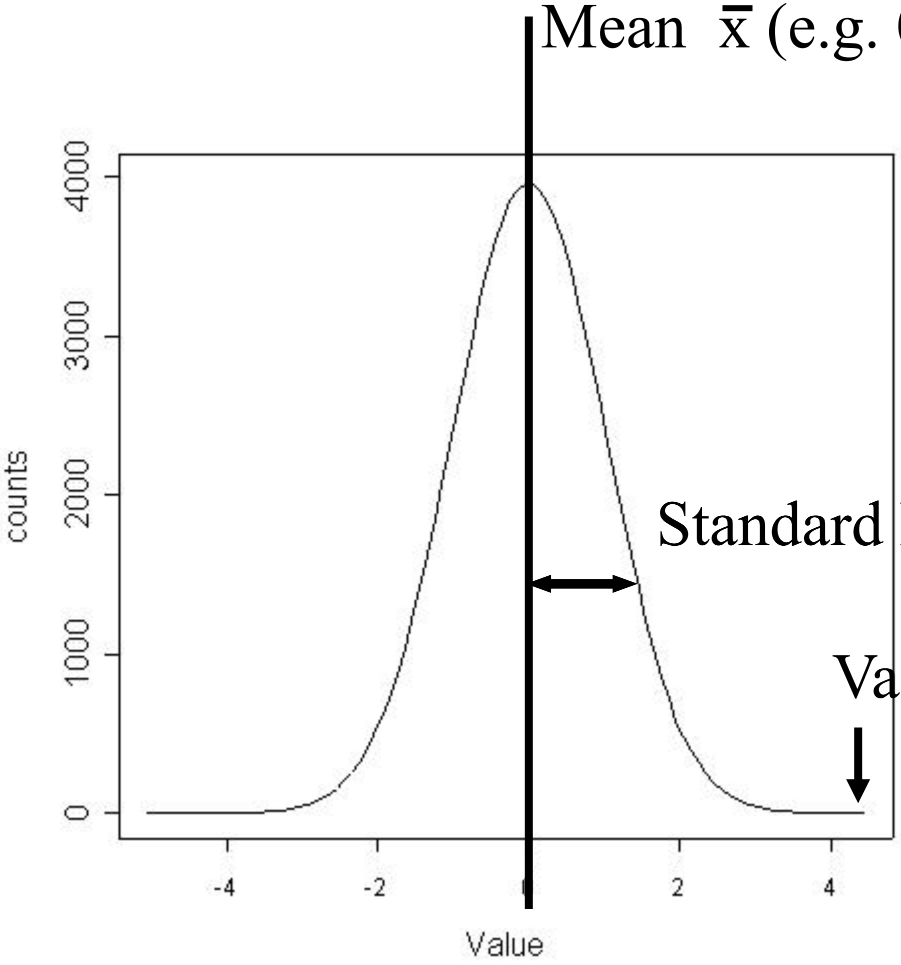
- Length-dependent
- Insensitive
- Dependent on the sequence alignment program and parameters
- Is a family of different scores…
  - Divide by length of shortest sequence
  - Divide by length of alignment
  - Divide by number of aligned positions etc.
    - See: Raghava, G.P.S. and Barton, G. J. Quantification of the variation in percentage identity for protein sequence alignments. BMC Bioinformatics. 2006 Sep 19;7:415.

Z-score compared to percentage identity

- Corrects for alignment length
- Is as sensitive as the alignment method
- Less sensitive to changes in the alignment method
- Only one way to calculate it

Z-score

- Align sequences and record score *S.*
- Shuffle order of amino acids in the sequences and re-align the pair. Record the score for this alignment, repeat 100 times.
- Calculate mean and Standard Deviation (sd) of shuffled sequence comparison scores.
- Z= (S-mean)/sd



Mean $\bar{x}$ (e.g. 0.0)

Standard Deviation σ (e.g. 1.8)

Value V (e.g. 4.3)

Z-score = (Value – Mean)/(Standard Deviation)

= $(V - \bar{x}) / \sigma$

e.g. = (4.3-0.0)/1.8 = 2.39

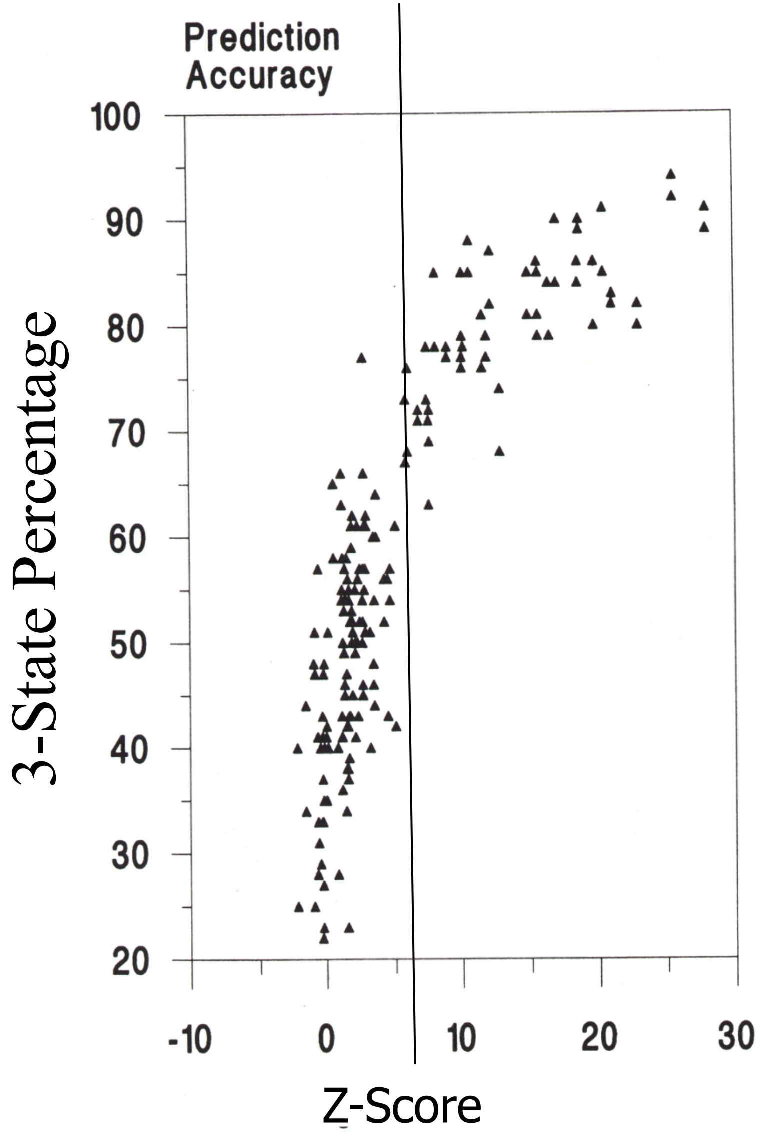


Fig. 2. The accuracy of secondary structure prediction by sequence alignment plotted against the alignment SD score to the homologous protein. One hundred and eighty-two predictions were made from pairwise alignment of the proteins in Table I.

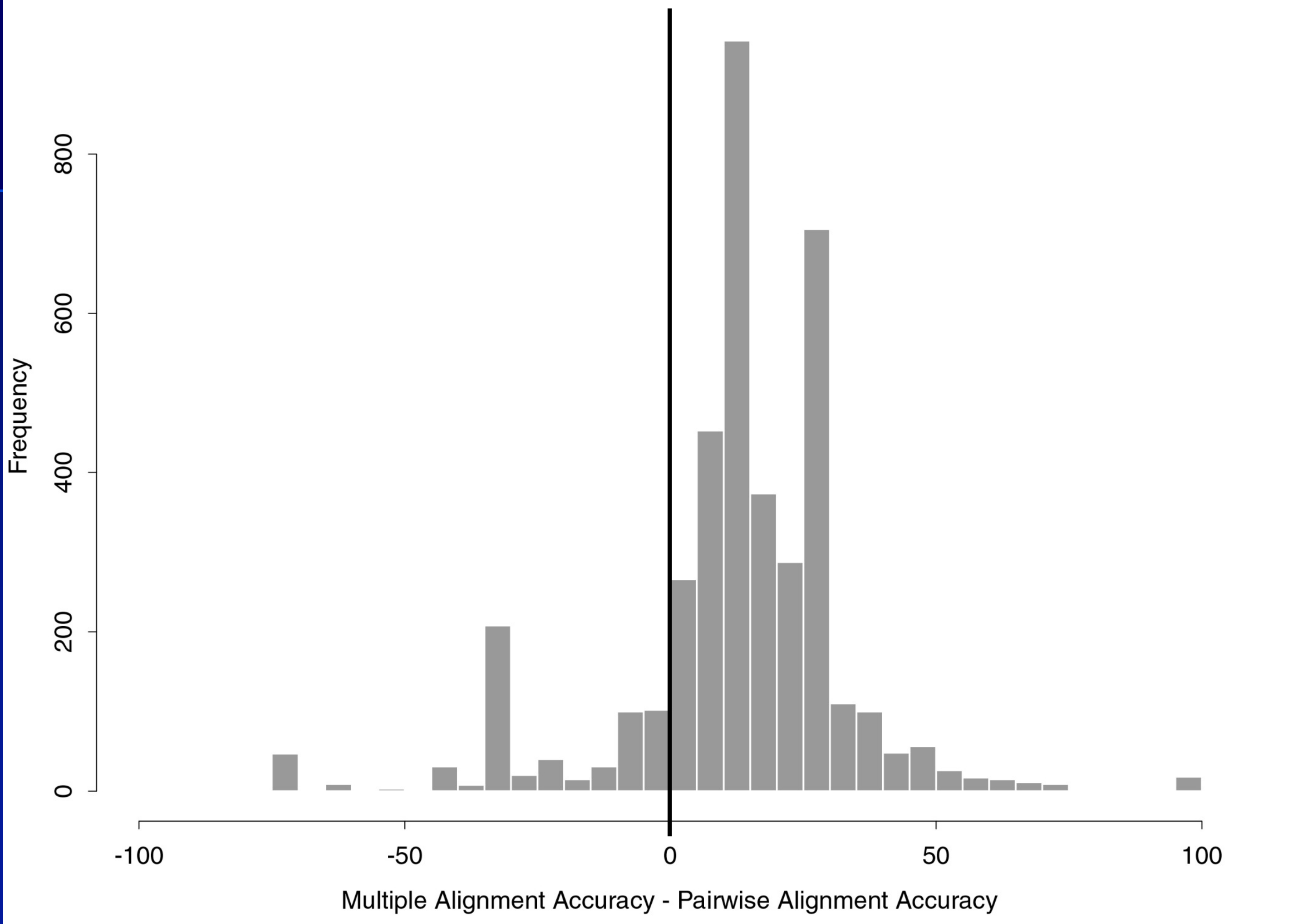
Alignment accuracy judged by agreement of secondary structure

No poor alignments for similarity > 6 sigma.

Boscott, P. E., Barton, G. J. and Richards, W. G. (1993), *Prot. Eng.,***6**, 261-266.

Alignment Accuracy Improves on Multiple Alignment

Improvement in Alignment Accuracy on Multiple Alignment



So, multiple alignments are on average more accurate than pair-wise alignments

Why is multiple alignment more accurate in general?

Single sequence

N Q L E V F M D G E L A . . .

physico-chemical properties of amino acids

Multiple sequences

```
N Q L E V F M D G E L E A . . .
N D E K V Y M E G D I Q V . . .
```

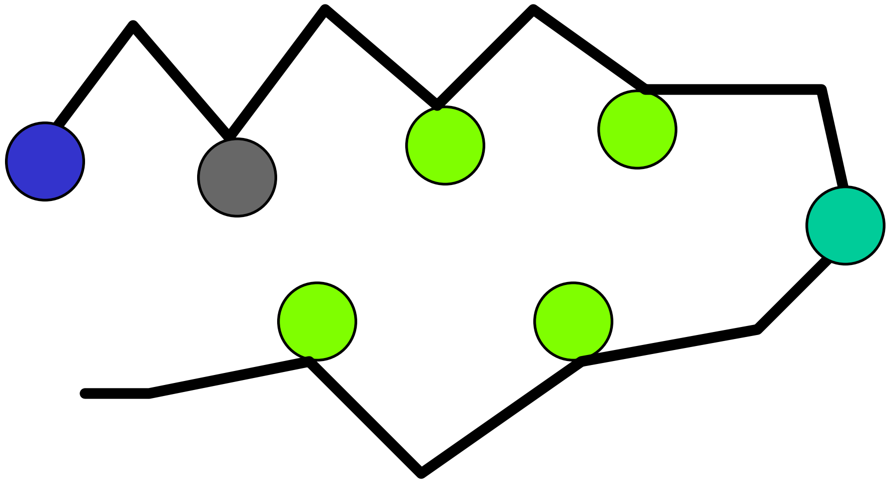
Multiple sequences

N	Q	L	E	**V**	F	**M**	D	**G**	E	**L**	E	**A**	. . .
N	D	E	K	**V**	Y	**M**	E	**G**	D	**I**	Q	**V**	. . .
N	S	S	Q	**V**	K	**I**	K	**G**	Q	**V**	D	**L**	. . .
N	N	T	N	**V**	A	**M**	R	**G**	K	**M**	N	**T**	. . .

conserved positions with
conserved hydrophobics

Multiple sequences

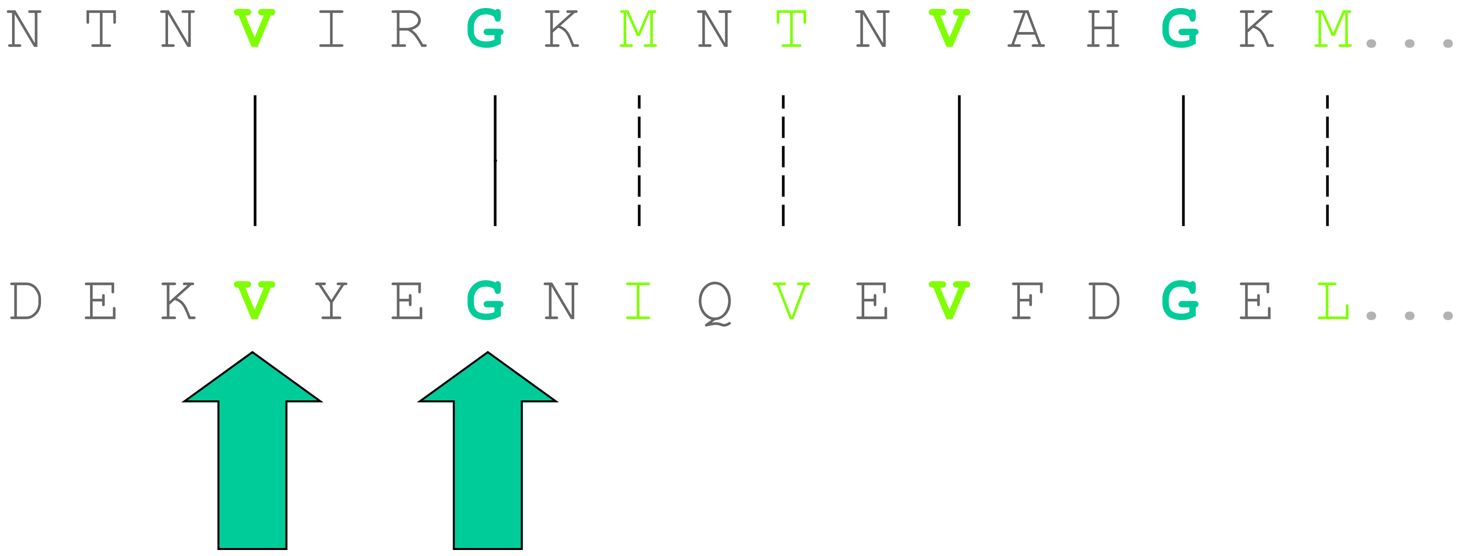
Think of a sequence on a structure



N	Q	L	E	**V**	F	**M**	D	**G**	E	**L**	E	**A**	. . .
N	D	E	K	**V**	Y	**M**	E	**G**	D	**I**	Q	**V**	. . .
N	S	S	Q	**V**	K	**I**	K	**G**	Q	**V**	D	**L**	. . .
N	N	T	N	**V**	A	**M**	R	**G**	K	**M**	N	**T**	. . .

Multiple sequences

help alignment itself



```
N T N V I R G K M N T N V A H G K M. . .

D E K V Y E G N I Q V E V F D G E L. . .
```

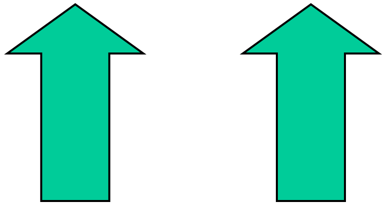
Watch these positions

Multiple sequences

help alignment itself (also pattern matching)

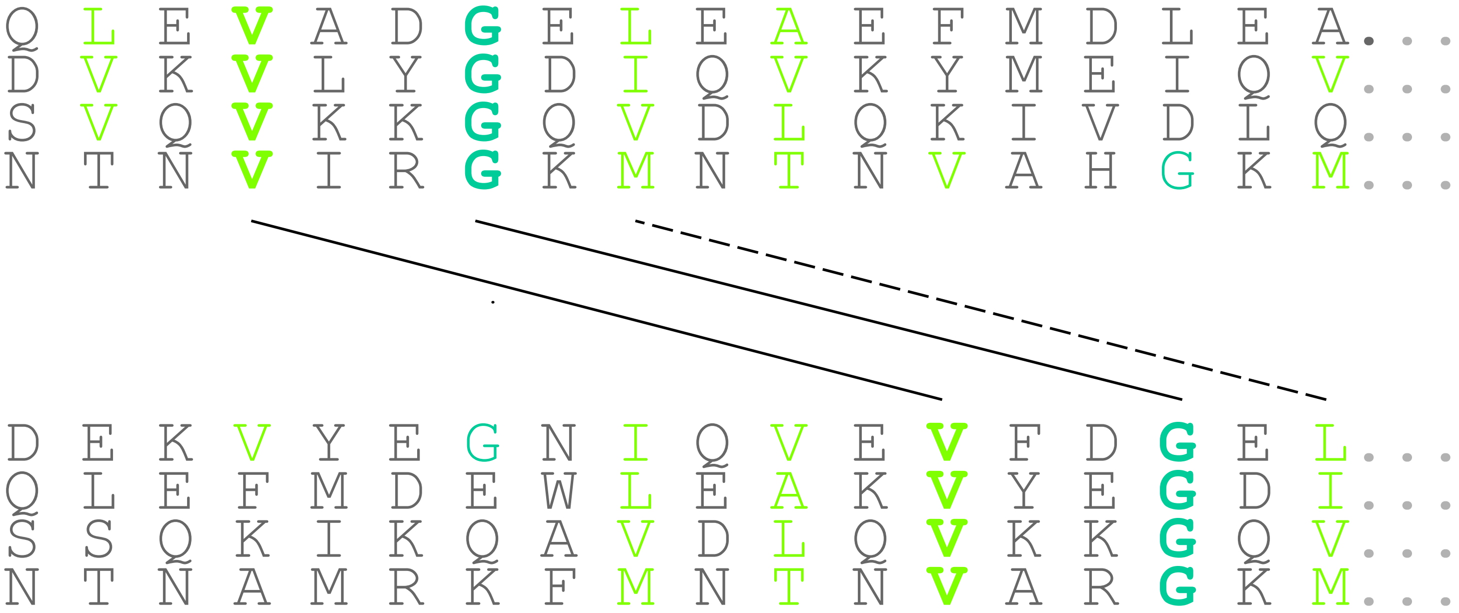
```
Q L E V A D G E L E A E F M D L E A...
D V K V L Y G D I Q V K Y M E I Q V...
S V Q V K K G Q V D L Q K I V D L Q...
N T N V I R G K M N T N V A H G K M...
```

```
D E K V Y E G N I Q V E V F D G E L...
Q L E F M D E W L E A K V Y E G D I...
S S Q K I K Q A V D L Q V K K G Q V...
N T N A M R K F M N T N V A R G K M...
```



Multiple sequences

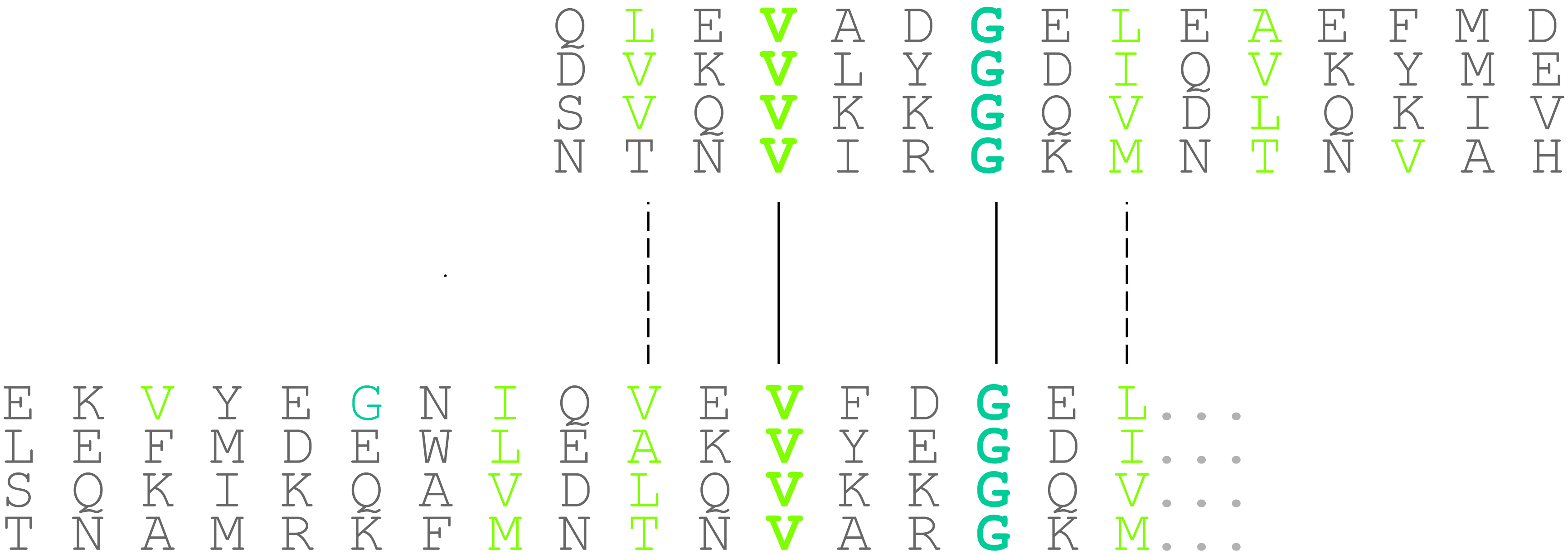
help alignment itself (also pattern matching)



```
Q L E V A D G E L E A E F M D L E A...
D V K V L Y G D I Q V K Y M E I Q V...
S V Q V K K G Q V D L Q K I V D L Q...
N T N V I R G K M N T N V A H G K M...

D E K V Y E G N I Q V E V F D G E L...
Q L E F M D E W L E A K V Y E G D I...
S S Q K I K Q A V D L Q V K K G Q V...
N T N A M R K F M N T N V A R G K M...
```

Multiple sequences

help alignment itself (also pattern matching)



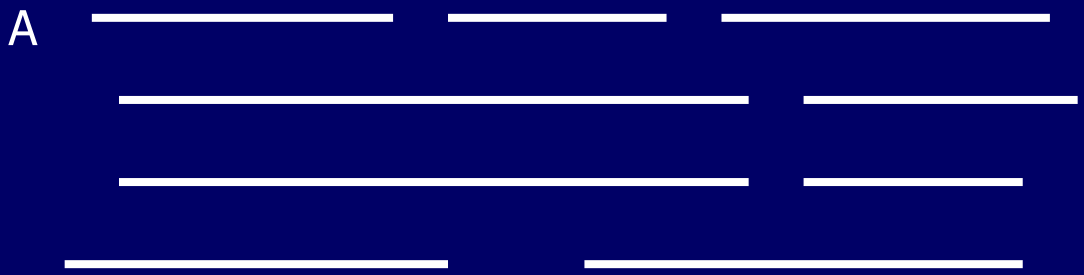
```
                Q L E V A D G E L E A E F M D
                D V K V L Y G D I Q V K Y M E
                S V Q V K K G Q V D L Q K I V
                N T N V I R G K M N T N V A H

E K V Y E G N I Q V E V F D G E L...
L E F M D E W L E A K V Y E G D I...
S Q K I K Q A V D L Q V K K G Q V...
T N A M R K F M N T N V A R G K M...
```

Multiple alignments for different purposes

- Obtain best “full” alignment of N sequences.
  - Good starting point for most purposes.

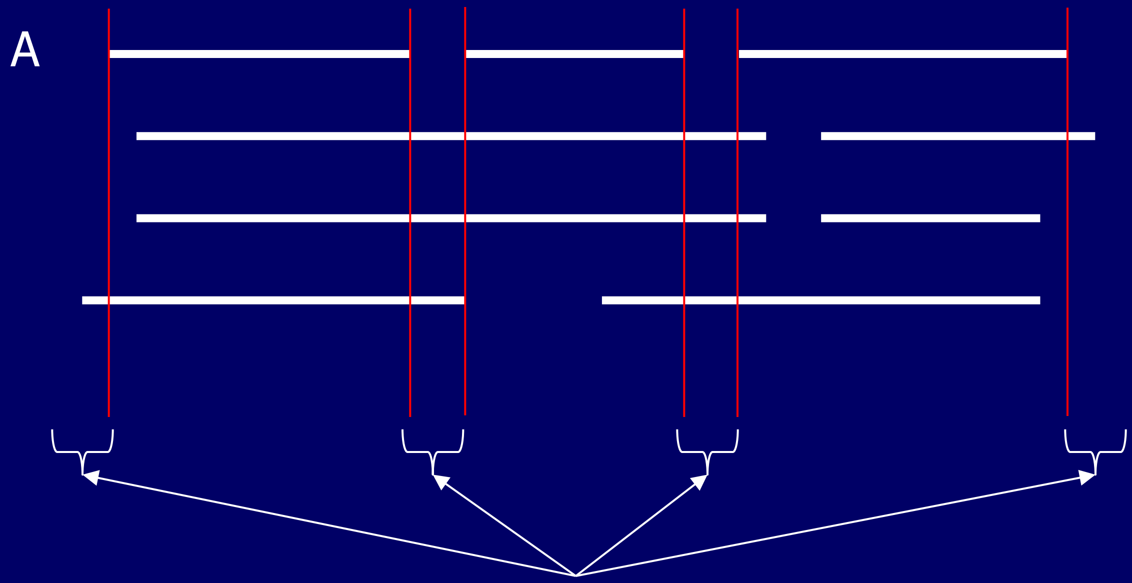
- Obtain best alignment relative to one sequence
  - Best when subsequent analysis is focused on the first sequence.
  - Best approach for iterative profile searching since it prevents the alignment length growing longer than the sequence.

Multiple alignments for different purposes



Normal Hierarchical alignment: Gaps appear in the first sequence if needed

Alignment specific to sequence A.



Alignment relative to first sequence only: Regions of second and subsequent sequences aligned with gaps in first Sequence are sometimes deleted. e.g. JPRED output. and some PSIBLAST output.

Regions deleted from alignment

Some uses of multiple Alignments

- Basis for sensitive profile searching of databases
- Identification of functional sites
- Phylogeny
- Presentation of sequence-related results
- Improved prediction of
  - Secondary structure
  - Disorder
  - Transmembrane regions
  - Almost any sequence-related property

Some uses of multiple Alignments

- Basis for sensitive profile searching of databases
- Identification of functional sites
- Phylogeny
- Presentation of sequence-related results
- Improved prediction of
  - Secondary structure
  - Disorder
  - Transmembrane regions
  - Almost any sequence-related property

Why are multiple alignments useful for prediction?

- Evolution highlights amino acids important to maintaining the structure and function of a protein
- This information can be captured by visual analysis, or better, by machine learning techniques such as Artificial Neural Networks.
- This is what Day 3 of this course is about!

Jalview – a tool with which to tackle many of these analyses

Also good for RNA and DNA

Jim Procter

Suzanne Duce

Mungo Carstairs

Tochukwu
(Charles)
Ofoegbu

Kira Mourao

Jalview

www.jalview.org
twitter:@jalview

First developed in 1996

elixir
UNITED KINGDOM

Tier 1 Resource

Jalview supported by wellcome trust

BBSRC bioscience for the future

THE BARTON GROUP

UNIVERSITY OF DUNDEE

Jalview