Multiple Sequence Alignments
A Brief Introduction

EMBL-Australia Masterclass on Protein Sequence Analysis

Sydney, Australia
Monday 21st October 2013

Aidan Budd
EMBL Heidelberg

Session Goal

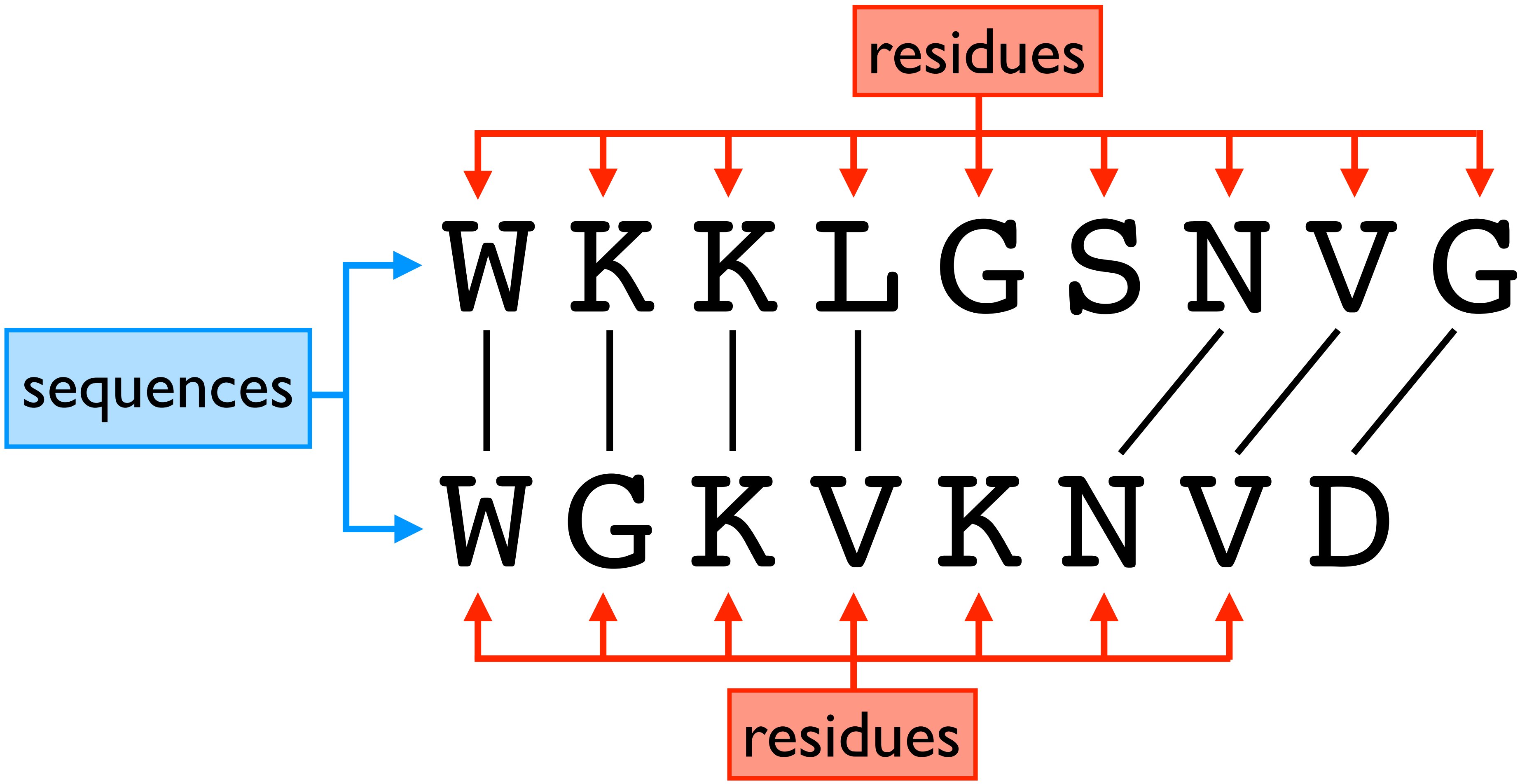
After attending today's session, we hope you will be better able to:

- build higher-quality/more appropriate MSAs for use in your own research/applications
- critically assess the quality of MSAs built by yourself and others

Why a Session on MSAs?

- Required for the development of almost all sequence analysis bioinformatics/tools
- MSAs take practice to interpret (and build) well
- Quality of downstream analysis/tools depends on quality of MSA

"Anatomy" of a Sequence Alignment

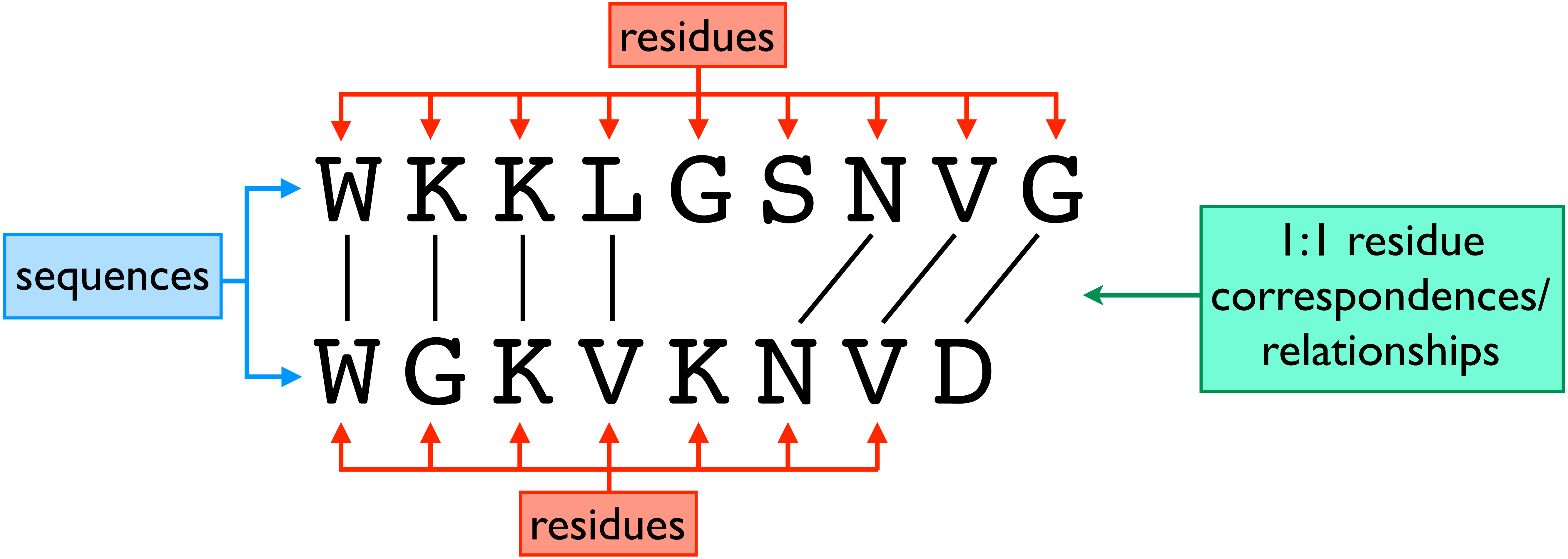


Residues:
Monomers within a polymer (polypeptide or polynucleotide) chain

Sequences:
List of residues in a polymer chain...
...listed in the same order they occur within the polymer

Aidan Budd, EMBL Heidelberg

"Anatomy" of a Sequence Alignment



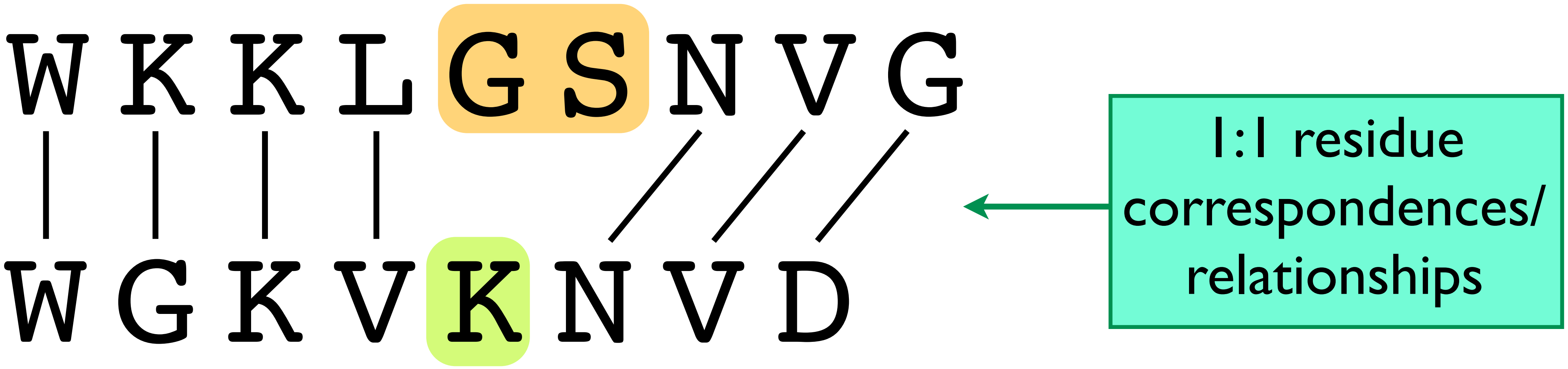
1:1 residue correspondences/relationships

Correspondences between

- a single residue in one sequence and
- a single residue in another sequence

Aidan Budd, EMBL Heidelberg

"Anatomy" of a Sequence Alignment



WKKLGSNVG

WGKVKNVD

1:1 residue correspondences/ relationships

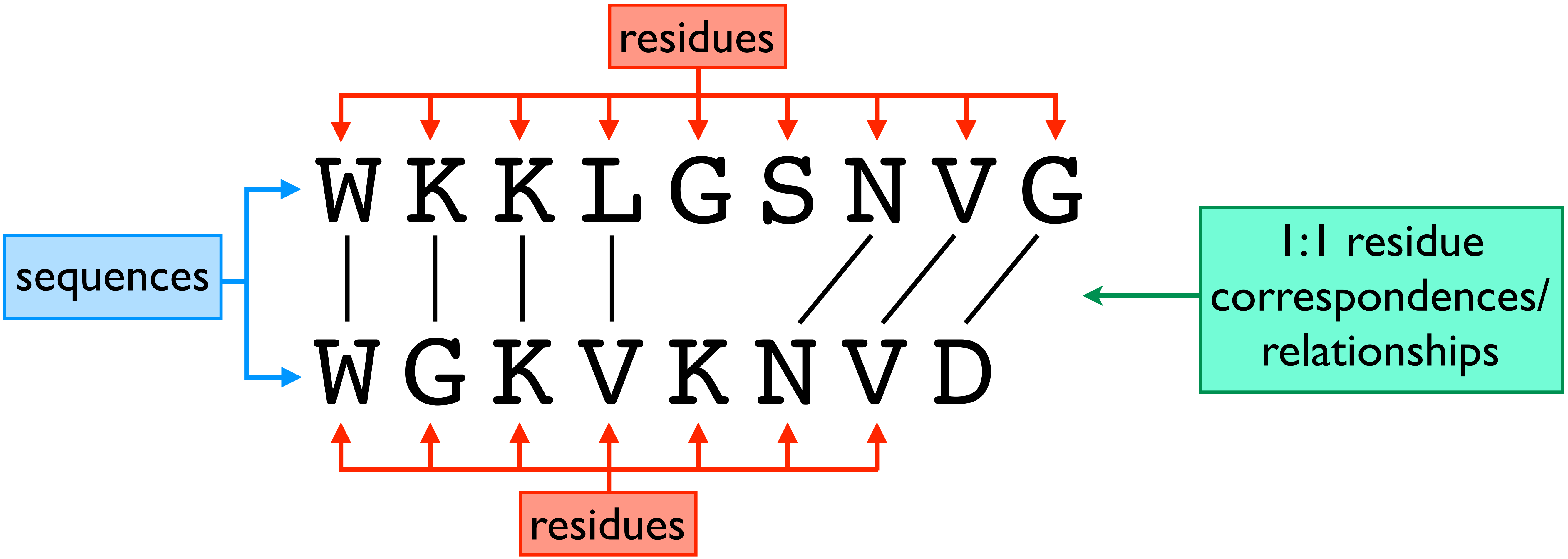
Residue has no equivalent in the top sequence

i.e. no residue in the top sequence has a 1:1 relationship with this residue

Could *perhaps* say there is a "1:2" relationship between this residue and these residues

However, alignments focus on 1:1 relationships

"Anatomy" of a Sequence Alignment



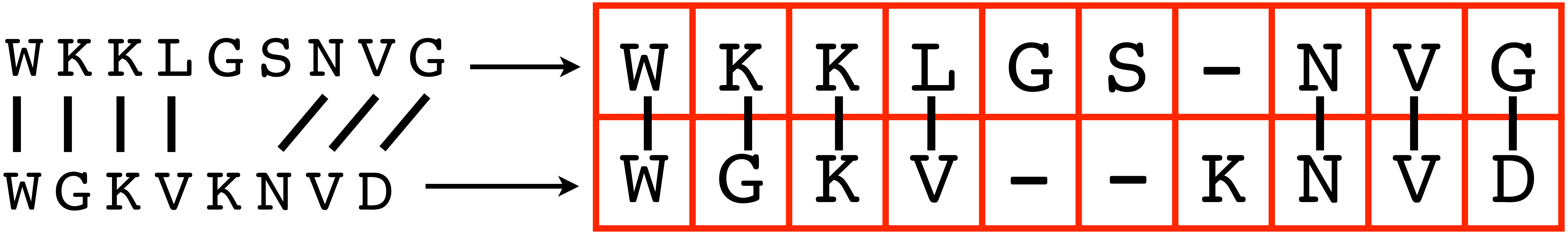
Sequence alignment

A comparison of the residues in two or more sequences...

... describing 1:1 correspondences/relationships/equivalences between residues in different sequences

Aidan Budd, EMBL Heidelberg

Sequence Alignment Within a Grid



W	K	K	L	G	S	-	N	V	G
W	G	K	V	-	-	K	N	V	D

Often represented using a grid/matrix :

One sequence per row

Residues in the same column are 'equivalent'

Gap characters (usually "-") indicate that the sequence contains no residues 'equivalent' to other residues in that column

Alternative Interpretations of MSAs (Evolutionary and Structural)

“Equivalence”/similarity of residues

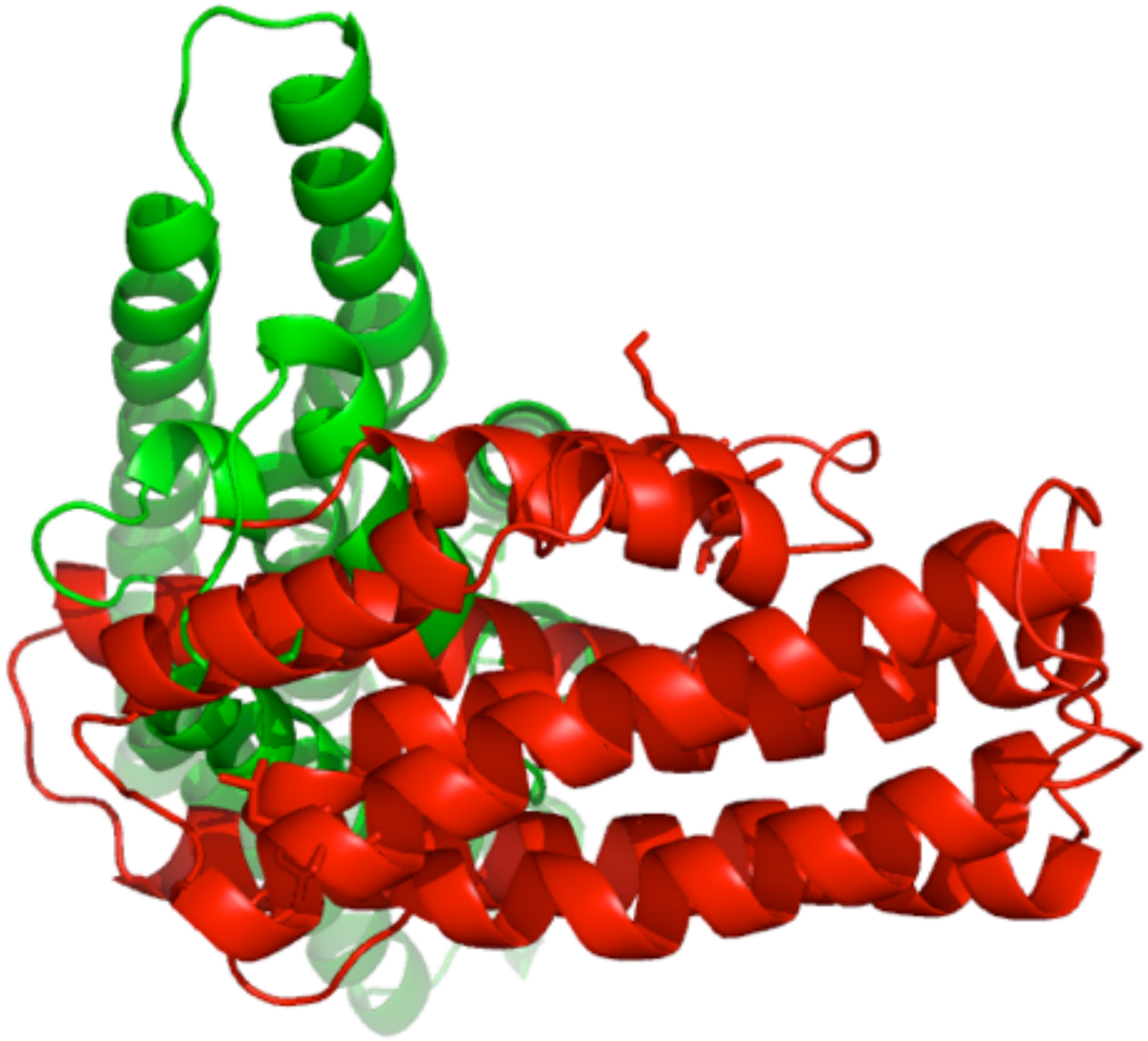
Residues in the same column either:

- Structurally equivalent/similar
- Evolutionary equivalent/related/homologous

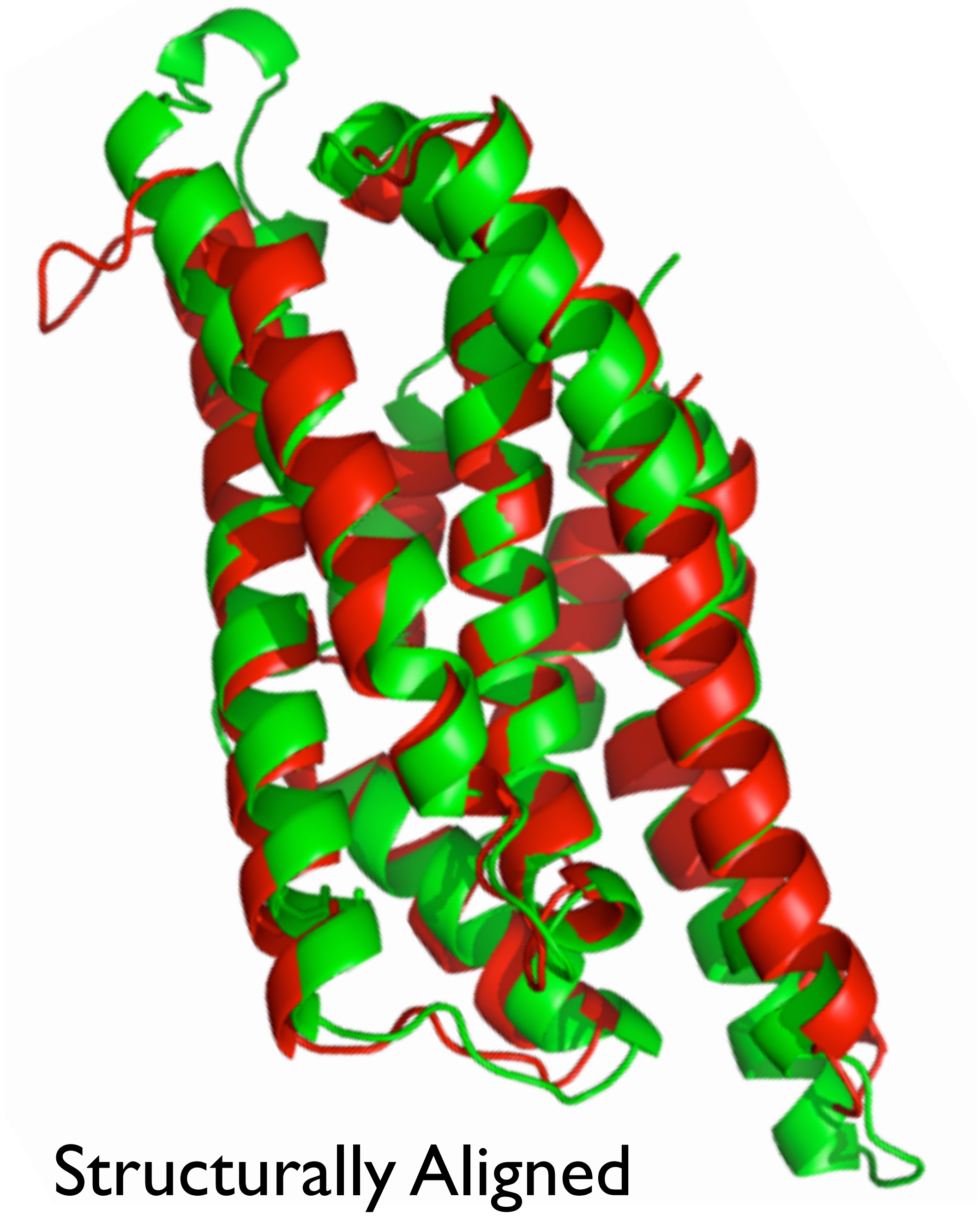
Different applications assume different types of equivalence

Different types of similarity not necessarily equivalent

Structural Similarity



Unaligned

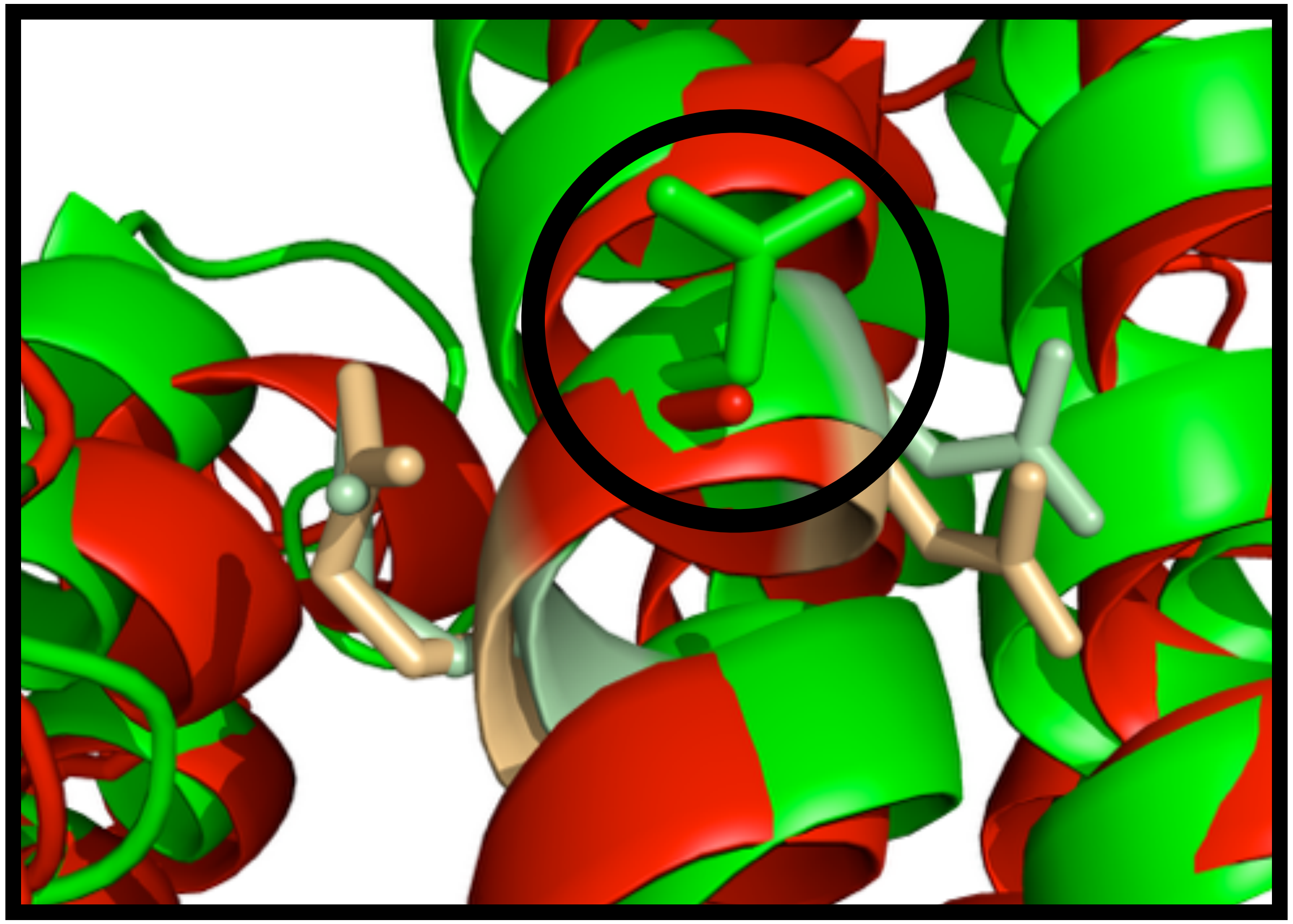


Structurally Aligned

Bacterial toxins 1ji6 and 1i5p

Aidan Budd, EMBL Heidelberg

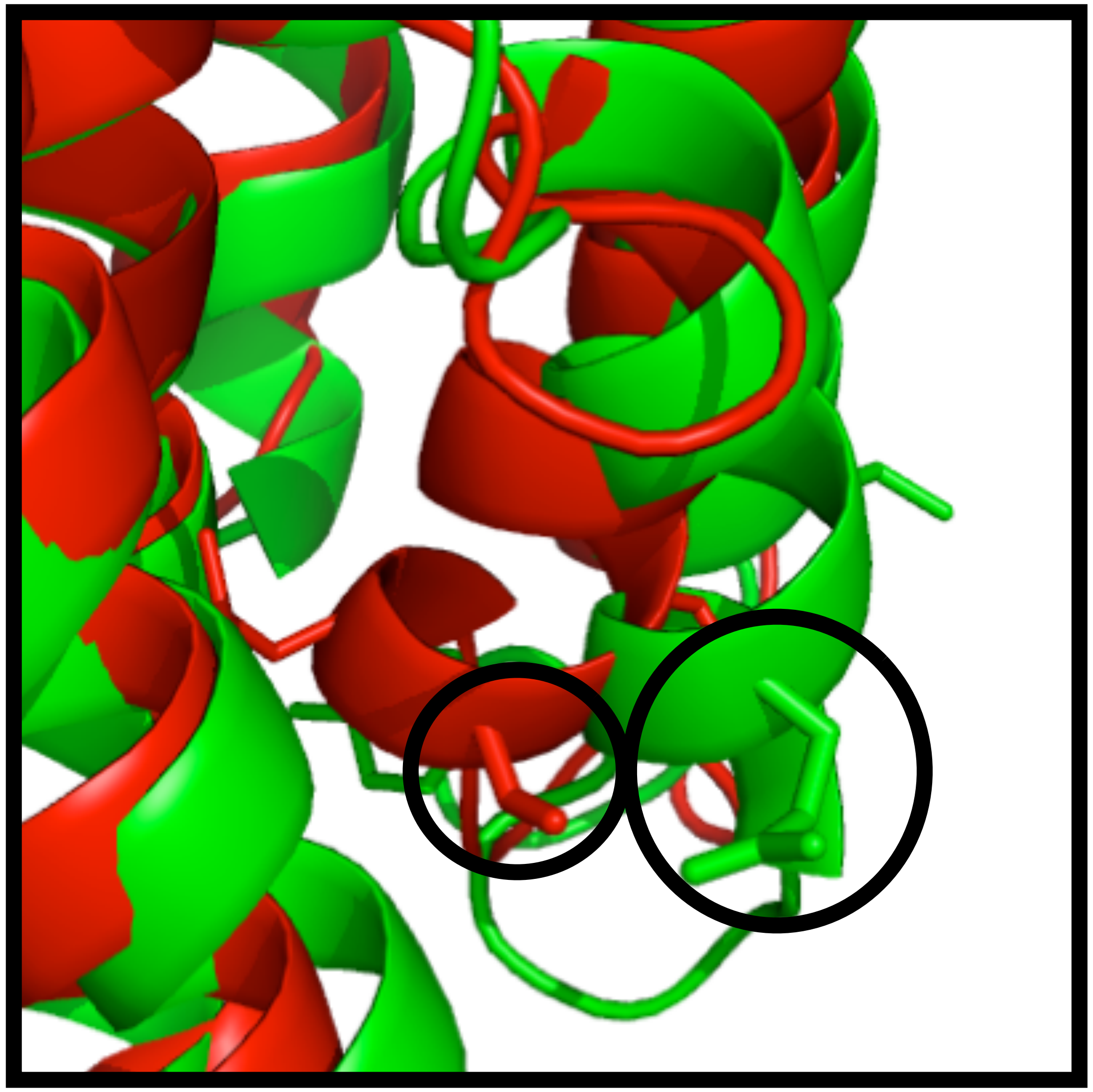
Structural Similarity

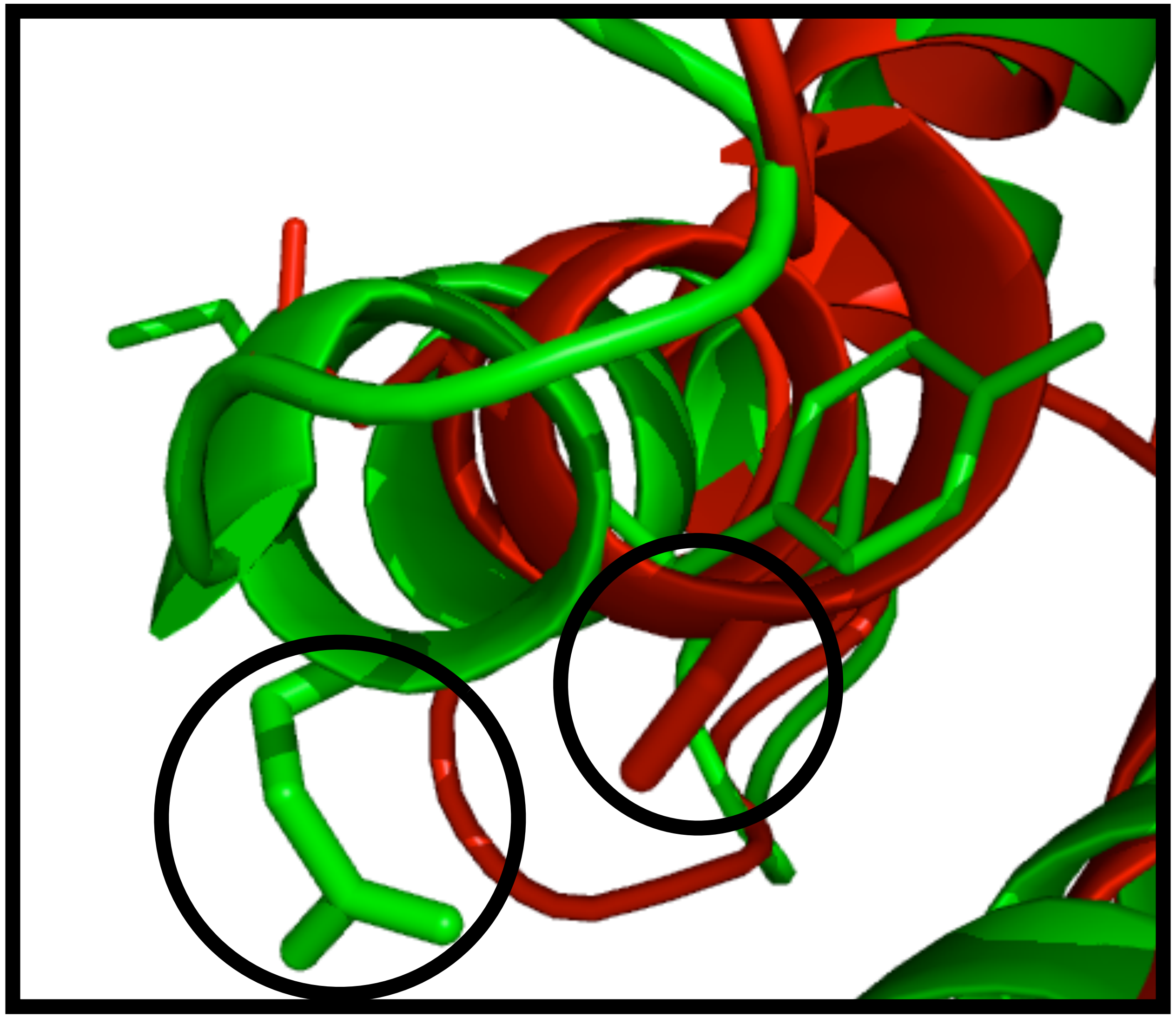


```
68  ELIGLQANIREFNQQVDNF
    1111111111111111111
70  ELQGLQNNFEDYVNALNSW
```

Residues with a similar structural context may lie almost on top of each other within a structural alignment. Clearly, the dark green and red side chains have more similar structural contexts than they do with the adjacent light-coloured side chains

Structural Similarity





```
Chain 1: 16 KVGSLIGKR---ILSELWGIIFPSGST
            111111111   11111111111 111
Chain 2: 16 VVGVPFAGALTSFYQSFLNTIWP-SDA
```

Aidan Budd, EMBL Heidelberg

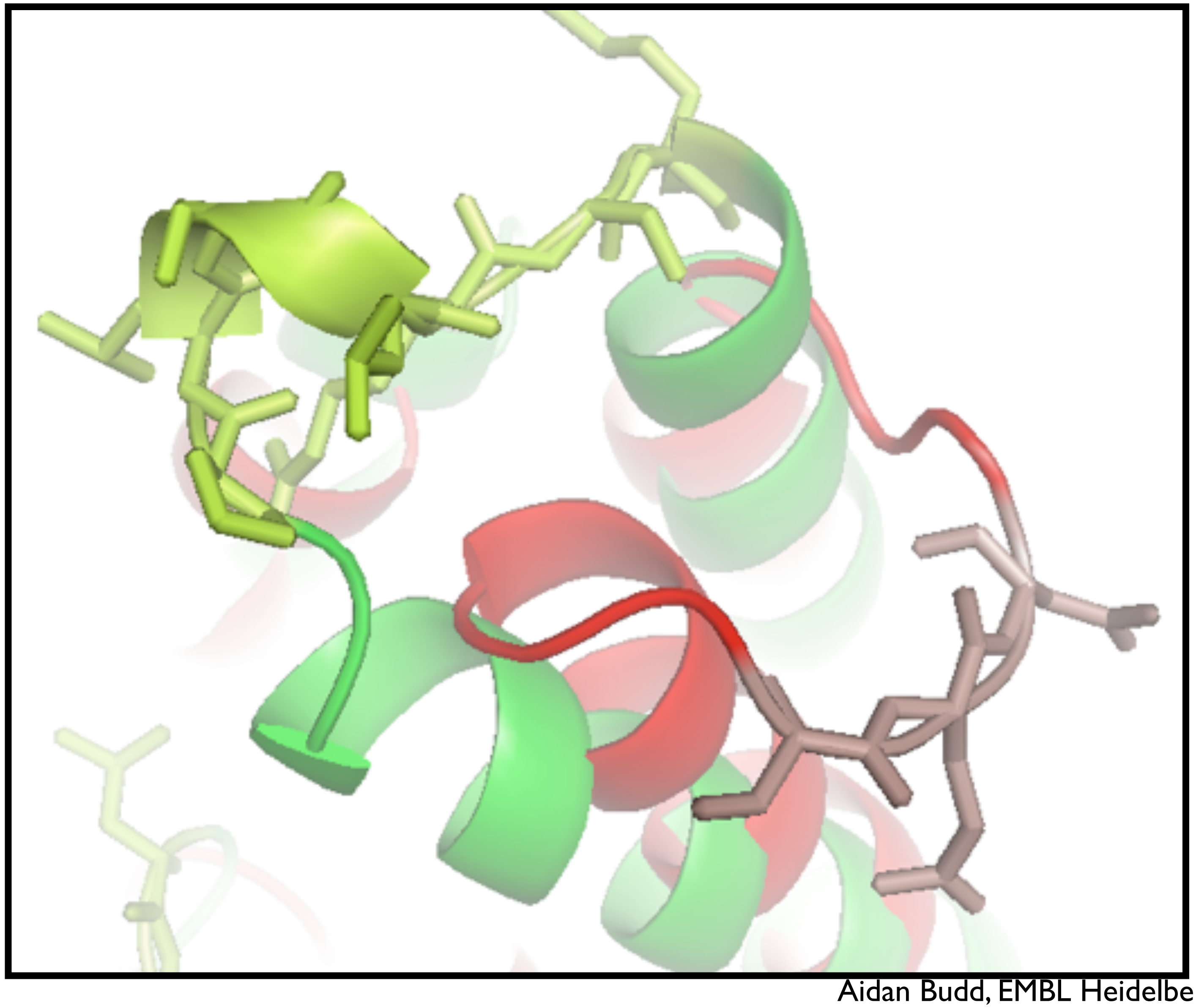
Structural equivalence

Some regions of the structures **do not have structurally equivalent residues** in the other structure

Alignment gaps are a sure sign of such residues

Placing such residues in the same column as residues from other sequences is a **misalignment** - to be avoided!

```
1i5p:    DNFLNPTQN----PVPLSITSSVN
         111111       11111111111
1ji6:    NSWKKTPLSLRSKRSQDRIRELFS
```

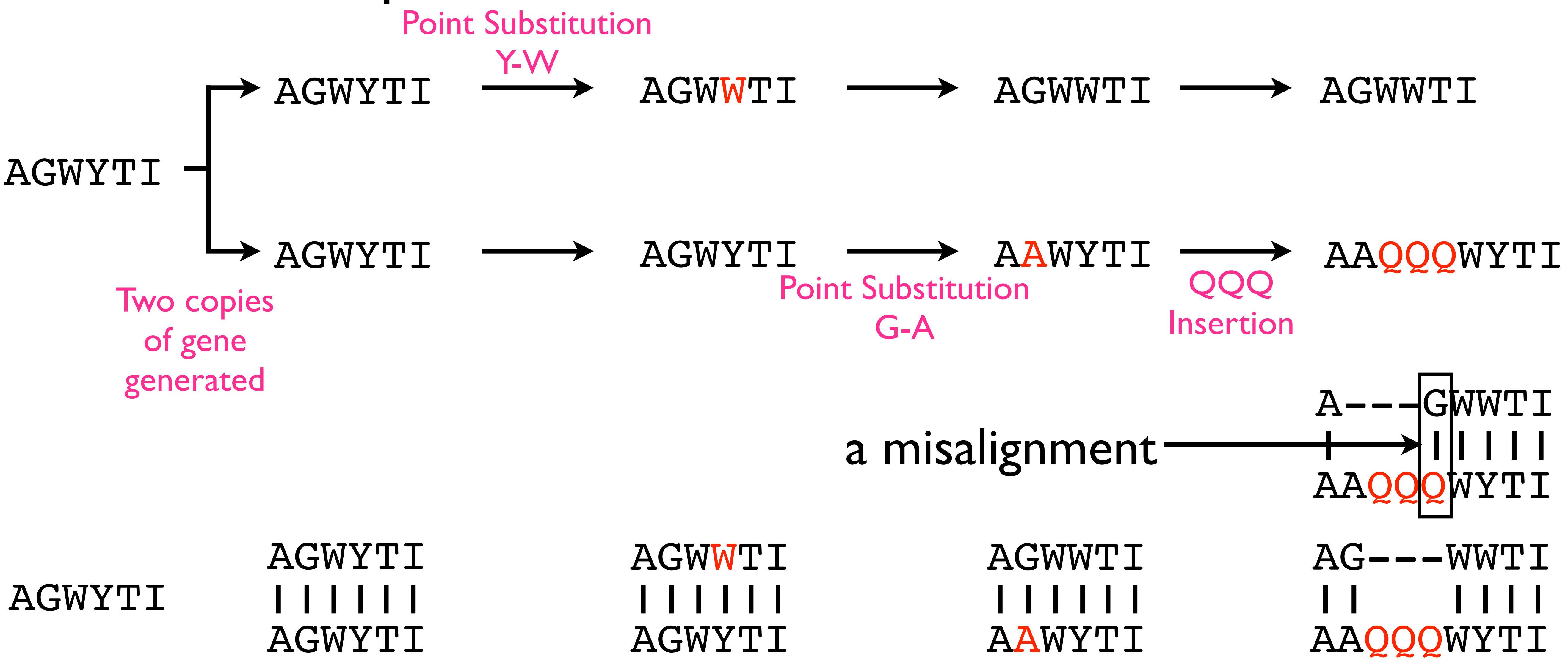


Aidan Budd, EMBL Heidelberg

Evolutionary "Equivalence”

Residues are "evolutionarily equivalent" when:

- they are derived from the same residue in an ancestral sequence
- the only mutations experienced during divergence from this ancestral residue were **point substitutions**



Point Substitution Y-W

AGWYTI → AGWYTI → AGWWTI → AGWWTI → AGWWTI

AGWYTI → AGWYTI → AGWYTI → AAWYTI → AAQQQWYTI

Two copies of gene generated

Point Substitution G-A

QQQ Insertion

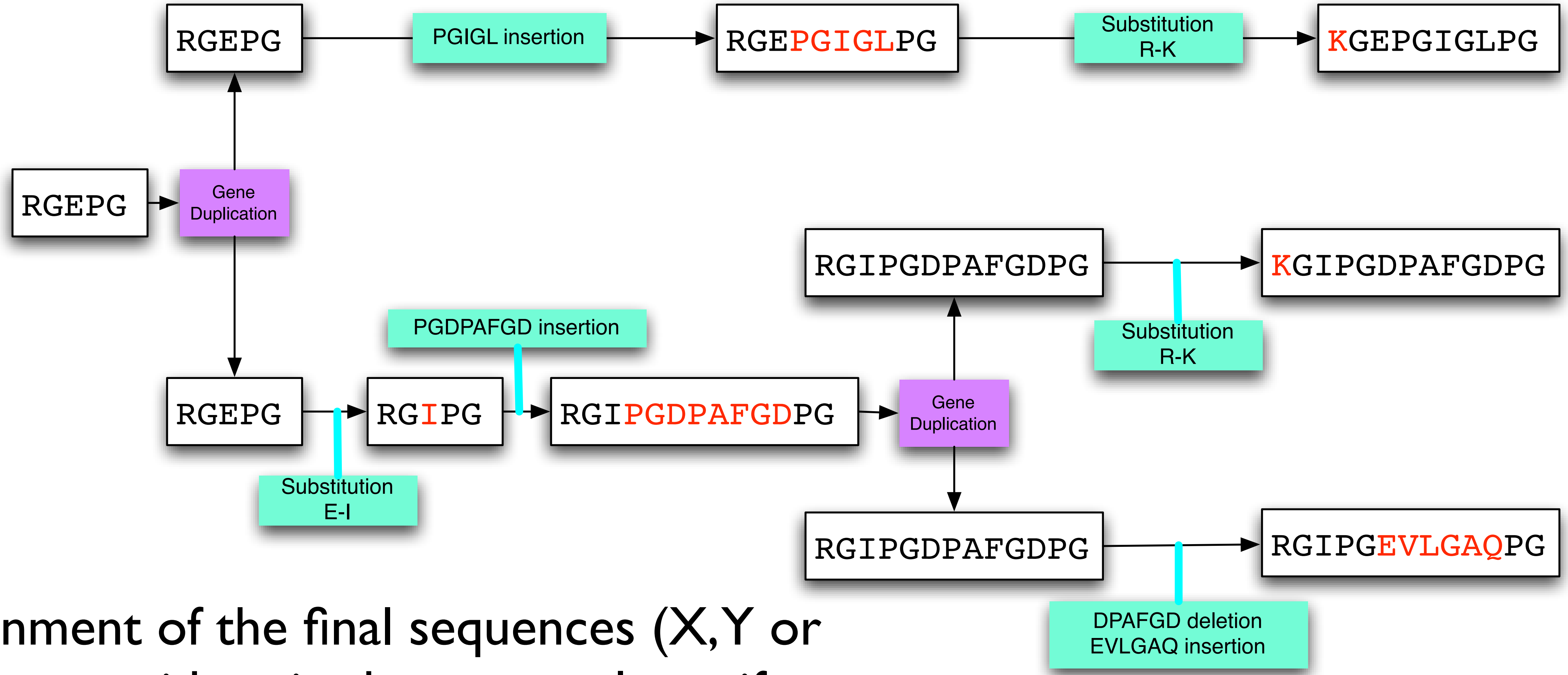
a misalignment

```
A---GWWTI
|    |||||
AAQQQWYTI
```

```
        AGWYTI   AGWWTI   AGWWTI   AG---WWTI
AGWYTI  ||||||   ||||||   ||||||   ||   ||||
        AGWYTI   AGWYTI   AAWYTI   AAQQQWYTI
```

Aidan Budd, EMBL Heidelberg

Quiz - Evolutionary Interpretation of Alignments



Which alignment of the final sequences (X,Y or Z) only places residues in the same column if they are related by substitution events?

X

```
KGEPG---IGLPG
KGIPGDPAFGDPG
RGIPGEVLGAQPG
```

Y

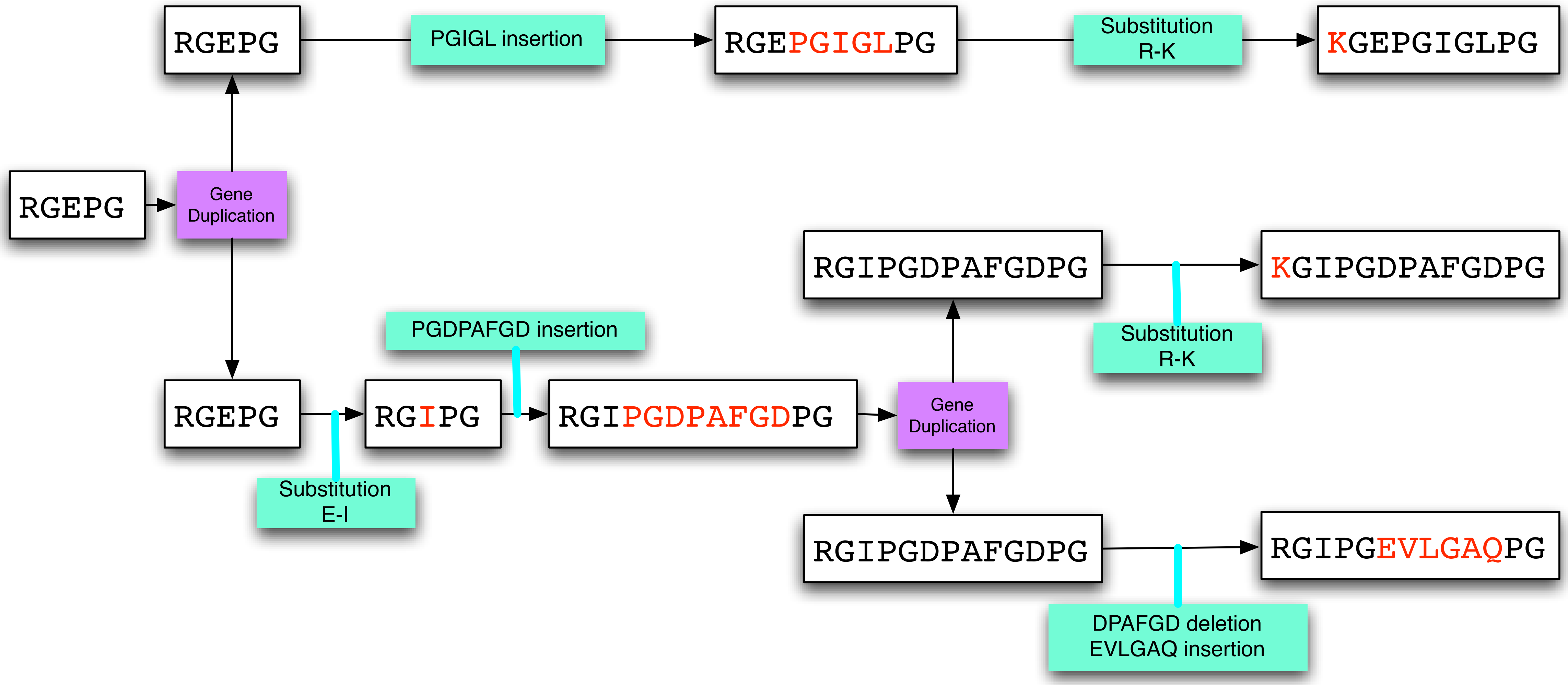
```
KGEPG------IGL------PG
KGIPG---------DPAFGDPG
RGIPGEVLGAQ---------PG
```

Z

```
KGE--------PGIGL------PG
KGIPG-----------DPAFGDPG
RGIPGEVLGAQ-----------PG
```

Aidan Budd, EMBL Heidelberg

Quiz - Evolutionary Interpretation of Alignments



"True" alignment given history described above

```
KGE--------PGIGL------PG
KGIPG----------DPAFGDPG
RGIPGEVLGAQ----------PG
```

PRANK

```
RGIPGEVLGAQPG
KGIPGDPAFGDPG
---KGEPGIGLPG
```

Quiz - Evolutionary Interpretation of Alignments

CLUSTALX

```
K---GEPGIGLPG
KGIPGDPAFGDPG
RGIPGEVLGAQPG
```

MAFFT

```
KGEPG---IGLPG
KGIPGDPAFGDPG
RGIPGEVLGAQPG
```

PRANK

```
RGIPGEVLGAQPG
KGIPGDPAFGDPG
---KGEPGIGLPG
```

Different automatic MSA software gives different results

All are different from the "true" alignment (assuming the scenario of transformation on the previous slide is true)...

... because that scenario is very unlikely under the models of evolutionary transformation incorporated within these tools

X

```
KGEPG---IGLPG
KGIPGDPAFGDPG
RGIPGEVLGAQPG
```

Y

```
KGEPG------IGL------PG
KGIPG---------DPAFGDPG
RGIPGEVLGAQ---------PG
```

Z

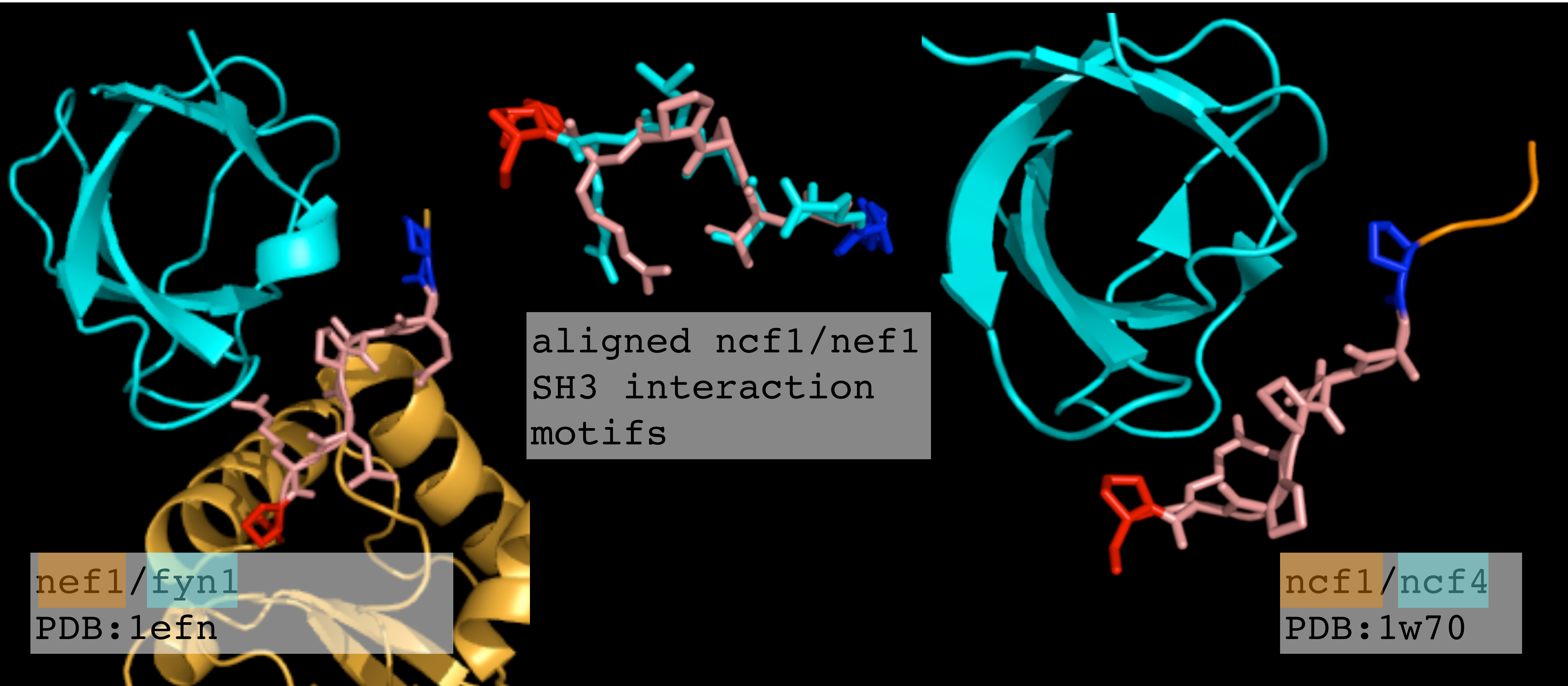
```
KGE--------PGIGL------PG
KGIPG-----------DPAFGDPG
RGIPGEVLGAQ-----------PG
```

Aidan Budd, EMBL Heidelberg

Non-Equivalence of Evolutionary and Structural Alignments

Structural equivalence without evolutionary equivalence

Structural alignment of SH3-interaction motifs from nef and ncf1



nef1/fyn1
PDB:1efn

aligned ncf1/nef1 SH3 interaction motifs

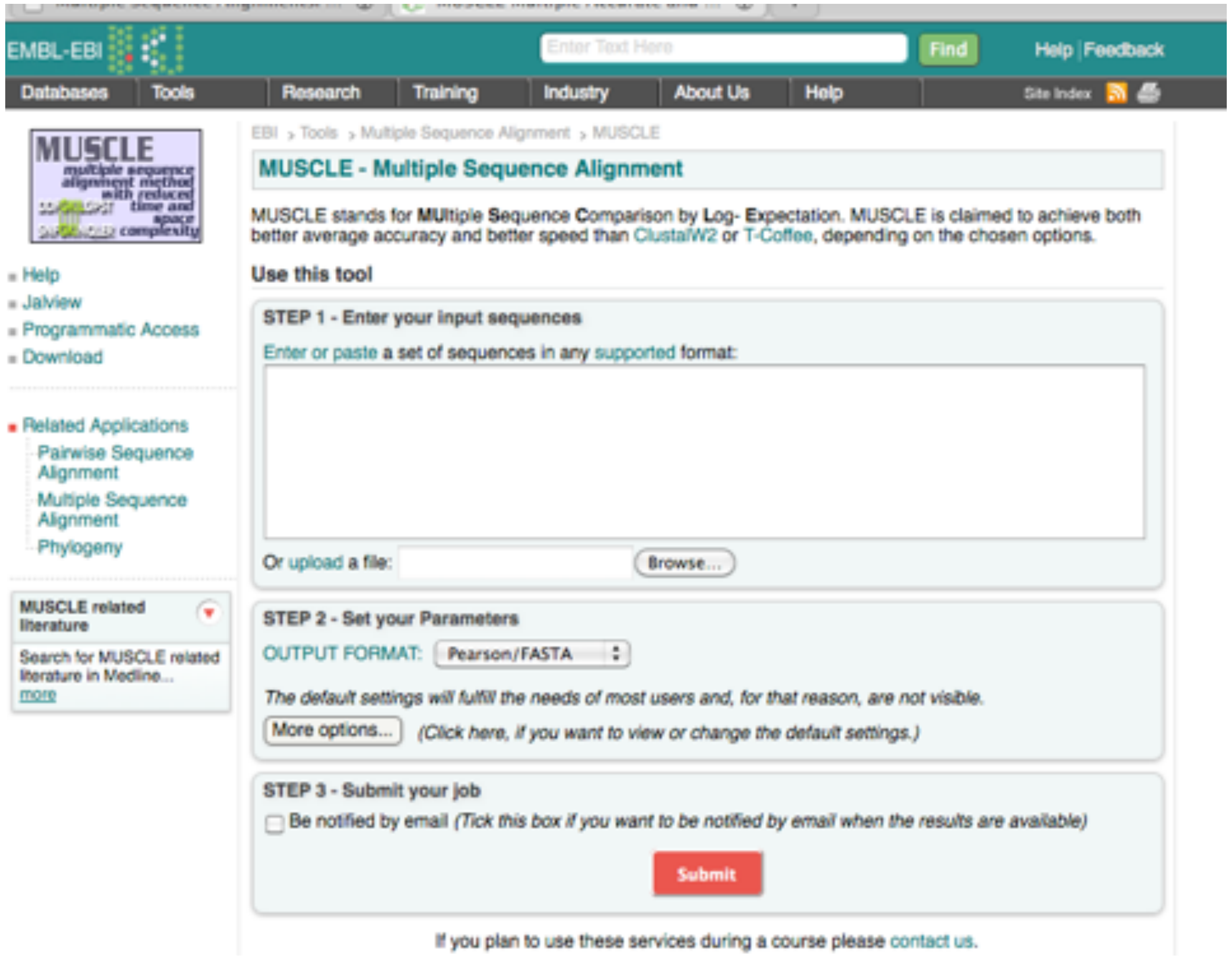
ncf1/ncf4
PDB:1w70

Building MSAs

Build an Automatic MSA

Load sequences into JalView, and with a few clicks you can automatically align a set of sequences

Or run an MSA tool at EBI



MUSCLE related literature

Search for MUSCLE related literature in Medline... more

MUSCLE - Multiple Sequence Alignment

MUSCLE stands for **MU**ltiple **S**equence **C**omparison by **L**og- **E**xpectation. MUSCLE is claimed to achieve both better average accuracy and better speed than ClustalW2 or T-Coffee, depending on the chosen options.

Use this tool

STEP 1 - Enter your input sequences

Enter or paste a set of sequences in any supported format:

Or upload a file:

STEP 2 - Set your Parameters

OUTPUT FORMAT: Pearson/FASTA

The default settings will fulfill the needs of most users and, for that reason, are not visible.

(Click here, if you want to view or change the default settings.)

STEP 3 - Submit your job

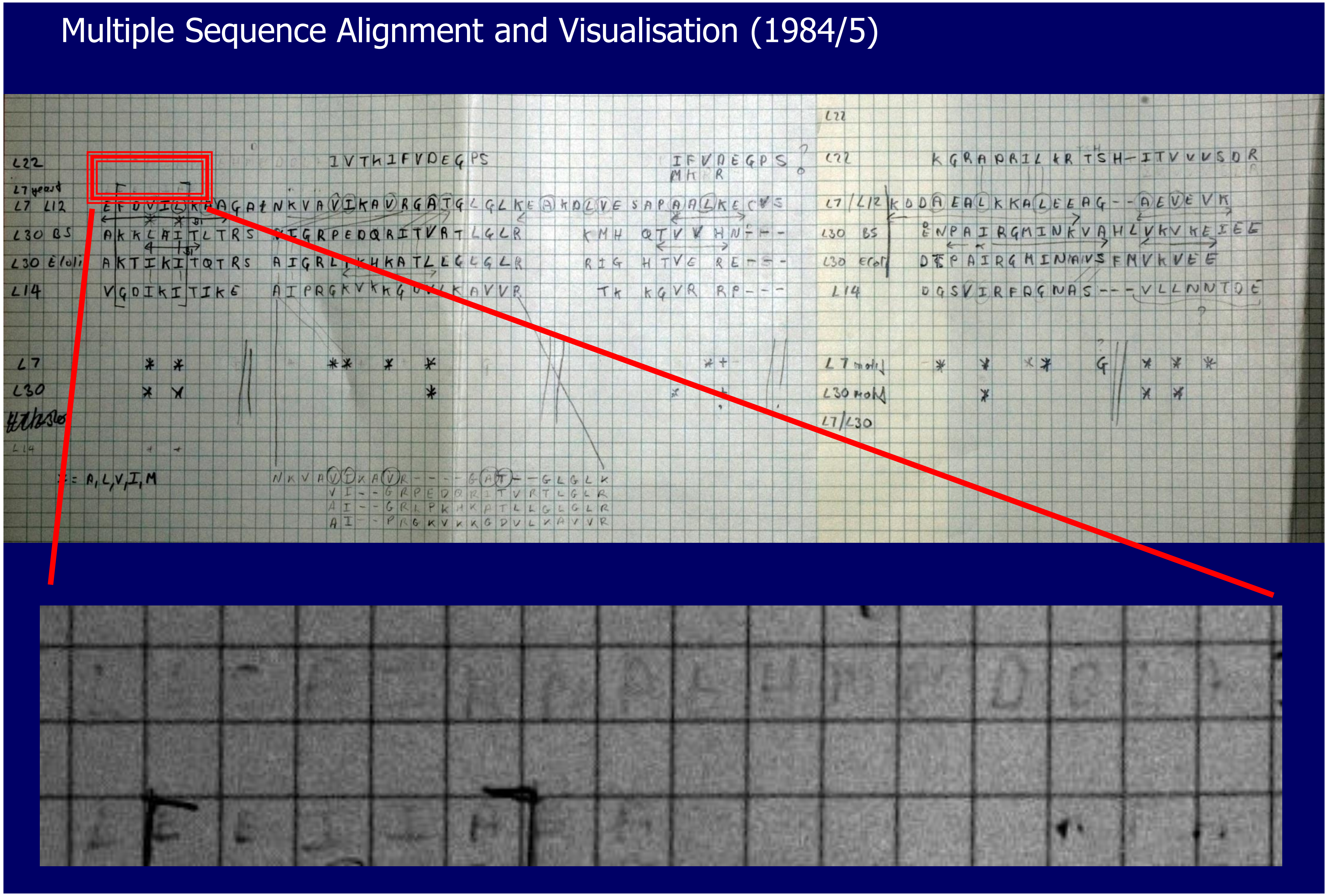
□ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

If you plan to use these services during a course please contact us.

http://www.ebi.ac.uk/Tools/msa/muscle/

Aidan Budd, EMBL Heidelberg

Build an MSA "Manually"



Multiple Sequence Alignment and Visualisation (1984/5)

L22 IVTkIFVDEGPS IFVDEGPS
MH R
L7 yeast
L7 L12 EFDVILKAAGAtNKVAVIKAVRGATGLGLKEAKDLVESAPAALKECVS
L30 BS AkKLAITLTRS ?IGRPEDQRITVRATLGLR KMH QTVV HN---
L30 E coli AKTIKITQTRS AIGRLPKHKATLLGLGLR RIG HTVE RE-S-
L14 VGDIKITIKE AIPRGKVKKGDVLKAVVR Tk KGVR RP---
L22 KGRAPRILkR TSH-ITVVVSDR
L7/L12 KDDAEALKKALEEAG--AEVEVK
L30 BS ENPAIRGMINKVAHLVKVKEIEE
L30 ECol DESPAIRGMINIAIVSFMVKVEE
L14 DGSVIRFAGNAS---VLLNNTDE
L7 * * ** * *
L30 * * *
L7 mod * * * * G * * *
L30 mod * * *
L7/L30
ɸ = A, L, V, I, M
NKVAVIKAVR----GAT--GLGLK
VI--GRPEDQRITVRTLGLR
AI--GRLPKHKATLLGLGLR
AI--PRGKVKKGDVLKAVVR

Courtesy of Geoff Barton, Dundee

JalView Demo and Exercises

- Loading sequences
- Changing the way the sequences are displayed
- Manual editing of alignments
- Adding/removing sequences to an alignment
- Exporting sequences/alignments from JalView for use in another application

JalView Demo and Exercises

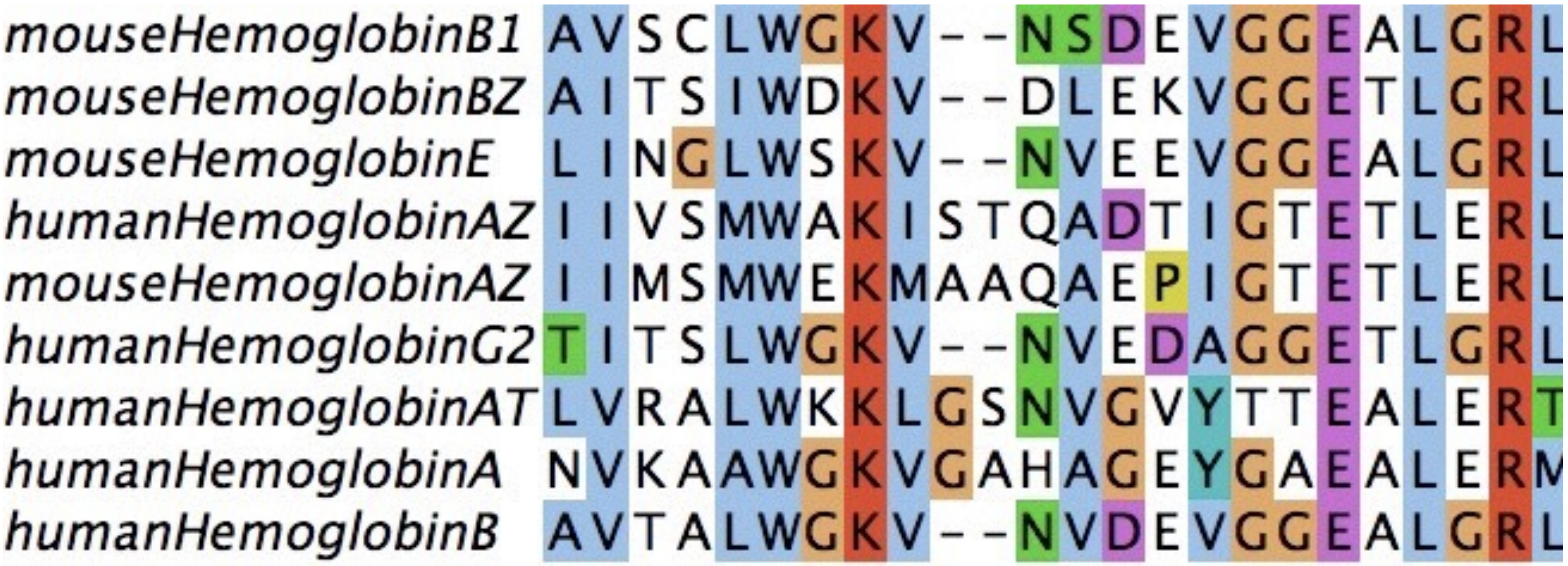
- a process of pattern-matching/identification
- we prefer alignments where many columns contain few differences/conservative changes
- more divergent sequences are harder to align than more similar sequences
    - for divergent sequences, there are many alternative alignments are similarly good/bad
    - for rather similar sequences, there is usually one/a few alignments we feel are clearly much better than the others
- Longer sequences take longer to align than short ones

JalView Demo and Exercises

- More sequences take longer to align than fewer sequences
- Repeats cause problems
- Different positions evolve differently
- At some level, the problem is "simple"
  - we just have to choose the right place to put the gaps!

another quiz on interpreting MSAs...

Quiz - Numbers of Insertions



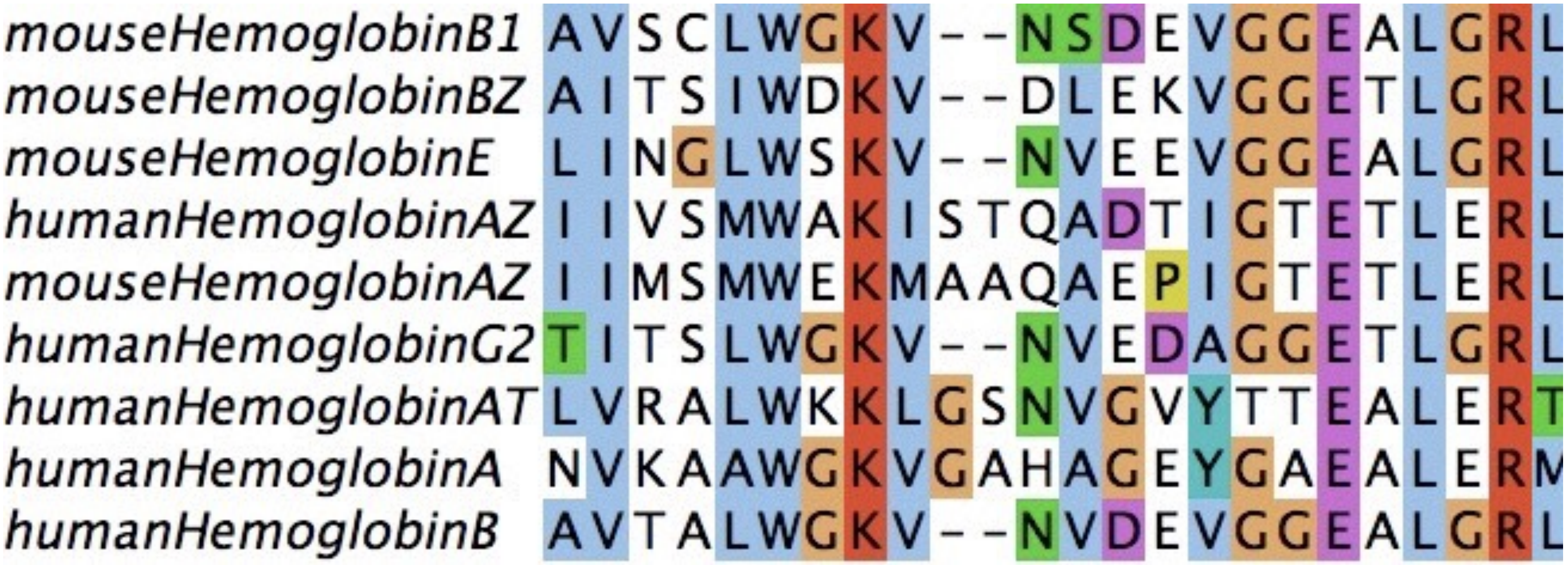
```
mouseHemoglobinB1   AVSCLWGKV--NSDEVGGEALGRL
mouseHemoglobinBZ   AITSIWDKV--DLEKVGGETLGRL
mouseHemoglobinE    LINGLWSKV--NVEEVGGEALGRL
humanHemoglobinAZ   IIVSMWAKISTQADTIGTETLERL
mouseHemoglobinAZ   IIMSMWEKMAAQAEPIGTETLERL
humanHemoglobinG2   TITSLWGKV--NVEDAGGETLGRL
humanHemoglobinAT   LVRALWKKLGSNVGVYTTEALERT
humanHemoglobinA    NVKAAWGKVGAHAGEYGAEALERM
humanHemoglobinB    AVTALWGKV--NVDEVGGEALGRL
```

The **minimum** number of insertion events required to account for the section of haemoglobin alignment shown above is?

(a) 2 (b) 1 (c) 0 (d) 3

Aidan Budd, EMBL Heidelberg

Quiz - Numbers of Insertions



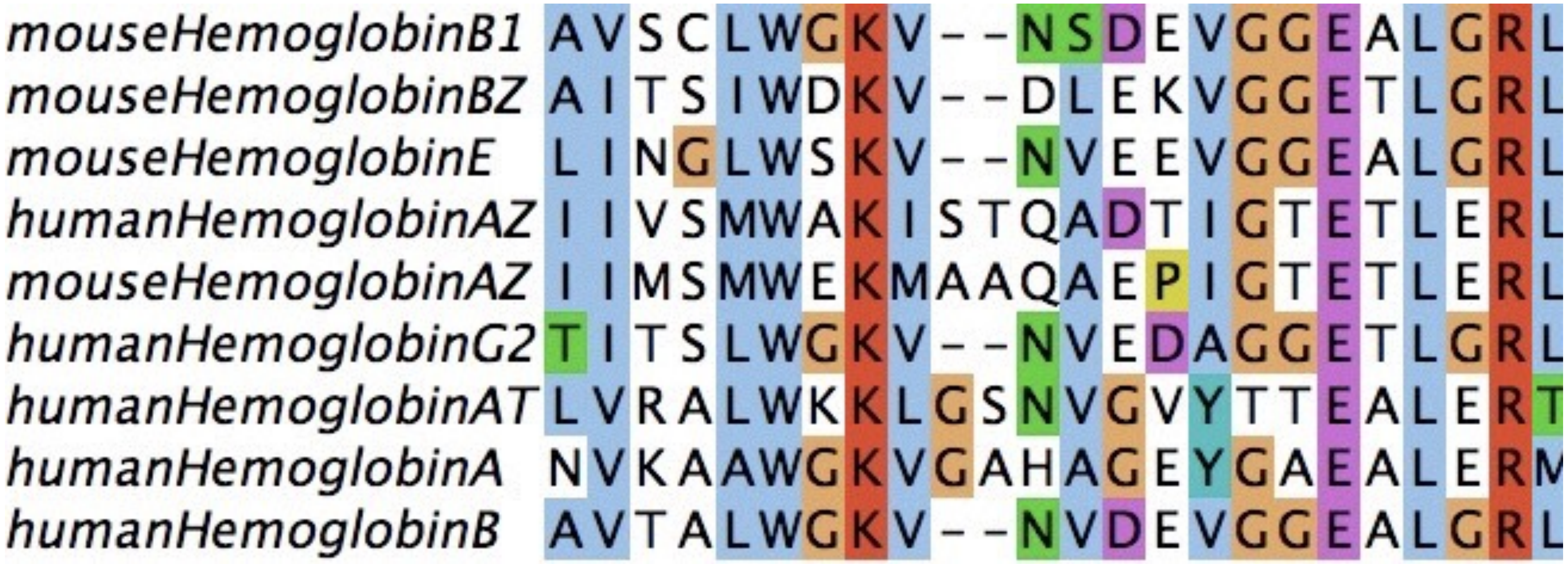
```
mouseHemoglobinB1   AVSCLWGKV--NSDEVGGEALGRL
mouseHemoglobinBZ   AITSIWDKV--DLEKVGGETLGRL
mouseHemoglobinE    LINGLWSKV--NVEEVGGEALGRL
humanHemoglobinAZ   IIVSMWAKISTQADTIGTETLERL
mouseHemoglobinAZ   IIMSMWEKMAAQAEPIGTETLERL
humanHemoglobinG2   TITSLWGKV--NVEDAGGETLGRL
humanHemoglobinAT   LVRALWKKLGSNVGVYTTEALERT
humanHemoglobinA    NVKAAWGKVGAHAGEYGAEALERM
humanHemoglobinB    AVTALWGKV--NVDEVGGEALGRL
```

The **minimum** number of insertion events required to account for the section of haemoglobin alignment shown above is?

If all sequences are the same length, we can explain their diversity without inferring ANY insertions or deletions

If and alignment contains sequences that are all either length **x** or **y**, then we can explain their diversity by inferring just one insertion or deletion

Quiz - Numbers of Insertions



```
mouseHemoglobinB1  AVSCLWGKV--NSDEVGGEALGRL
mouseHemoglobinBZ  AITSIWDKV--DLEKVGGETLGRL
mouseHemoglobinE   LINGLWSKV--NVEEVGGEALGRL
humanHemoglobinAZ  IIVSMWAKISTQADTIGTETLERL
mouseHemoglobinAZ  IIMSMWEKMAAQAEPIGTETLERL
humanHemoglobinG2  TITSLWGKV--NVEDAGGETLGRL
humanHemoglobinAT  LVRALWKKLGSNVGVYTTEALERT
humanHemoglobinA   NVKAAWGKVGAHAGEYGAEALERM
humanHemoglobinB   AVTALWGKV--NVDEVGGEALGRL
```

The **minimum** number of insertion events required to account for the section of haemoglobin alignment shown above is?

We can ALWAYS explain observed sequence length diversity with:

- 0 insertions (all length variation due to deletion)
- 0 deletions (all length variation due to insertion)
- a combination of insertions and deletions

Perhaps we should instead focus on inferring the **most likely** scenario?

(Although if this is not particularly relevant for our analysis, perhaps we should focus instead on something completely different!)