Introduction to Computational Biology; An Evolutionary Approach: Multiple Sequence Alignment

Bernhard Haubold & Thomas Wiehe

Outline

Multiple Sequence Alignment

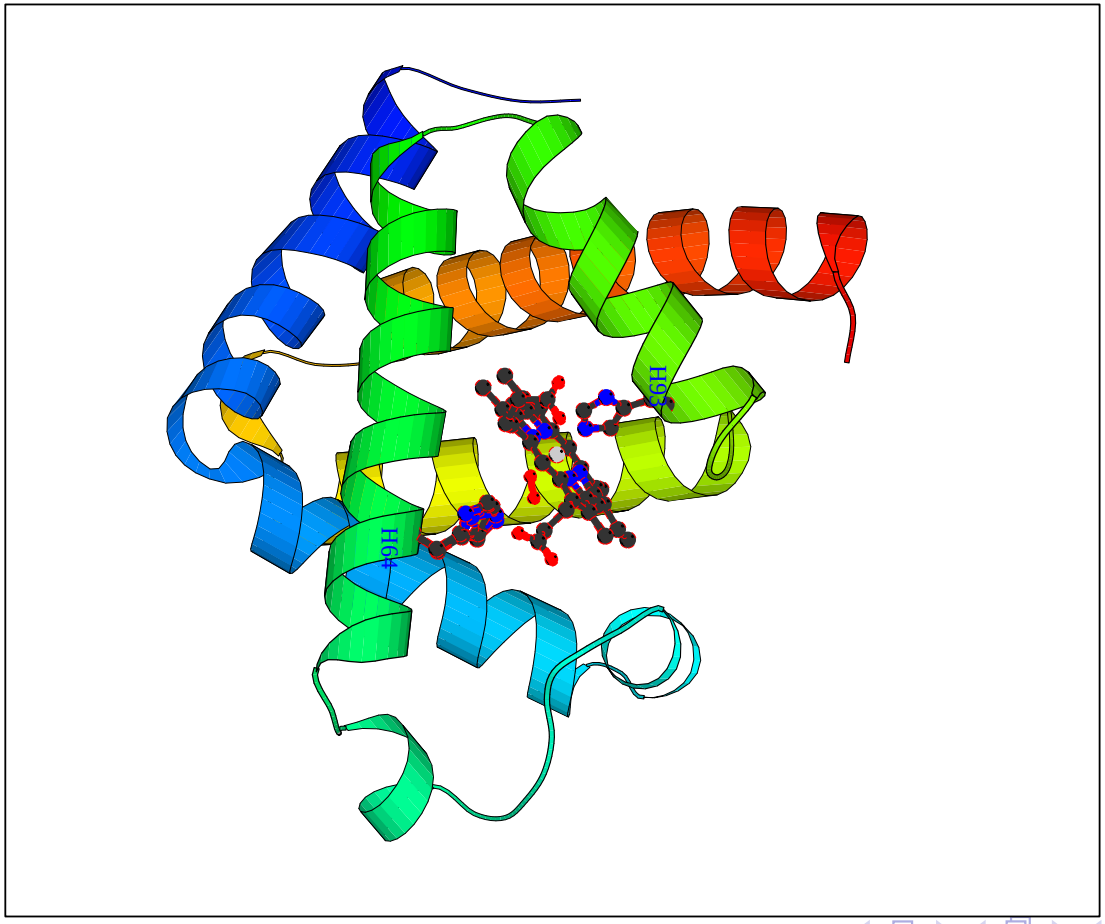
```
human β     --------VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQR    40
horse β     --------VQLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQR    40
human α     ---------VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKT    41
horse α     ---------VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKT    41
lamprey g   PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQE    50
whale m     ---------VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLE    41
lupine l    --------GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKD    42
consensus             !*  **  * * ! **     * * *** * *   ! *

                                  Conserved Histidine
                                          ↓
human β     FFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDN-----LKGTF    85
horse β     FFDSFGDLSNPGAVMGNPKVKAHGKKVLHSFGEGVHHLDN-----LKGTF    85
human α     YFPHF-DLS-----HGSAQVKGHGKKVADALTNAVAHVDD-----MPNAL    80
horse α     YFPHF-DLS-----HGSAQVKAHGKKVGDALTLAVGHLDD-----LPGAL    80
lamprey g   FFPKFKGLTTADQLKKSADVRWHAERIINAVNDAVASMDDT--EKMSMKL    98
whale m     KFDRFKHLKTEAEMKASEDLKKHGVTVLTALGAILKKKGH-----HEAEL    86
lupine l    LFSFLKGTSEVP--QNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATL    90
consensus    !  * ***      **  ***!****  *   ** ***         *

            Conserved Histidine
                    ↓
human β     ATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVA   135
horse β     AALSELHCDKLHVDPENFRLLGNVLVVVLARHFGKDFTPELQASYQKVVA   135
human α     SALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLA   130
horse α     SNLSDLHAHKLRVDPVNFKLLSHCLLSTLAVHLPNDFTPAVHASLDKFLS   130
lamprey g   RDLSGKHAKSFQVDPQYFKVLAAVIADTVAAG---------DAGFEKLMS   139
whale m     KPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALE   136
lupine l    KNLGSVHVSKG-VADAHFPVVKEAILKTIKEVVGAKWSEELNSAWTIAYD   139
consensus     !* *!* ** *** ** **    *  *** *     ***    *   *

human β     GVANALAHKYH------   146
horse β     GVANALAHKYH------   146
human α     SVSTVLTSKYR------   141
horse α     SVSTVLTSKYR------   141
lamprey g   MICILLRSAY-------   149
whale m     LFRKDIAAKYKELGYQG   153
lupine l    ELAIVIKKEMNDAA---   153
consensus    *   *   **
```

Oxy-Myoglobin



Sum-of-Pairs Score

- A: length of alignment
- n: number of sequences
- $s(S_j[i], S_k[i])$: score of position i of sequence j aligned with position i of sequence k; 2 gaps $\equiv 0$

$$M = \sum_{i=1}^{|\mathrm{A}|} \sum_{j}^{n} \sum_{k<j} s(S_j[i], S_k[i])$$

Path in Hyperlattice

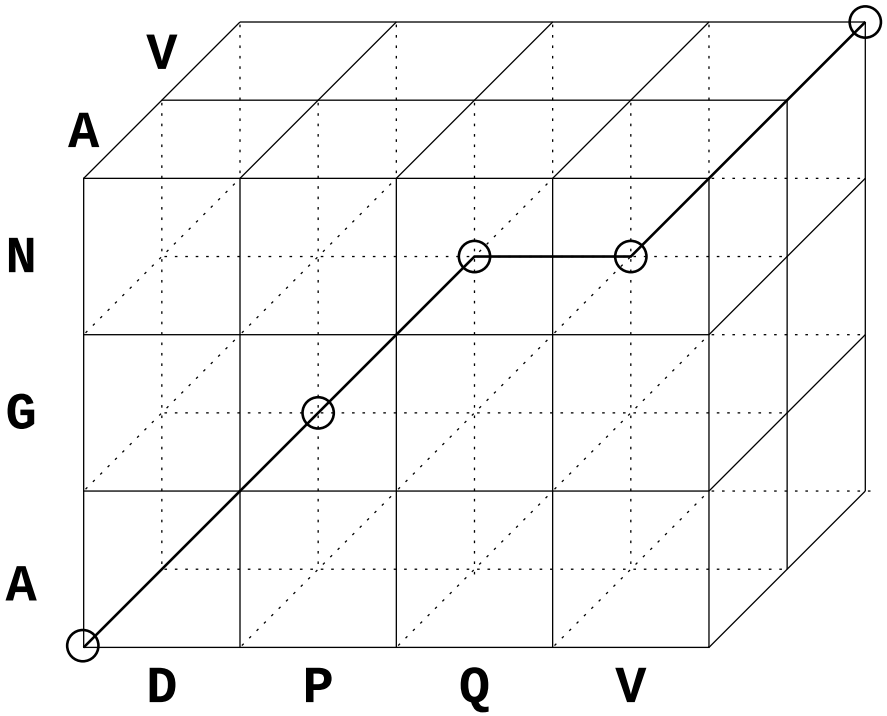
$$(0,0,0,0,0,0,0) \rightarrow (0,0,0,0,1,0,0) \rightarrow \ldots \rightarrow (0,0,0,0,8,0,0)$$
$$\rightarrow (1,1,0,0,9,0,1) \rightarrow (2,2,1,1,10,1,2) \rightarrow \ldots \rightarrow$$
$$(147,147,141,141,149,153,153)$$

Number of Vertices

$$\prod_{i=1}^{n}(|S_i| + 1)$$

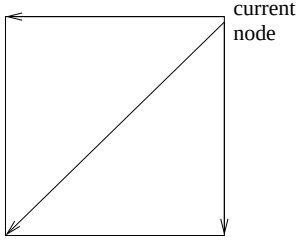
Exercise: Compute number of vertices in hyperlattice for aligning the seven globin sequences.

Multiple Sequence Alignment by Dynamic Programming

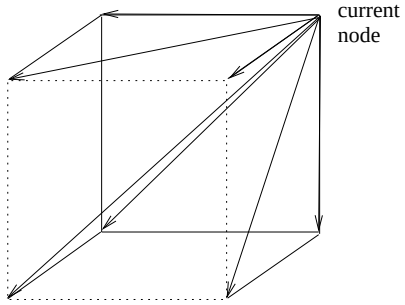



```
DPQV
AG-N
A--V
```

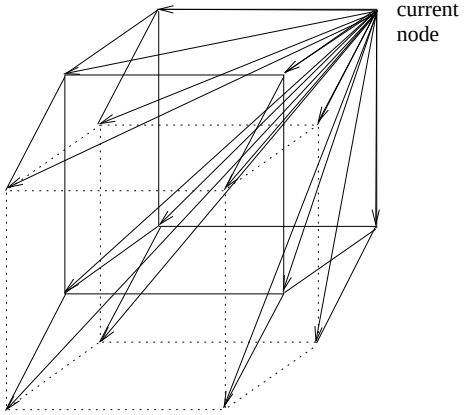
Nodes in Alignment Hyperlattice



2 sequences: 3 vertices to visit



3 sequences: 7 vertices to visit



4 sequences: 15 vertices to visit

Optimal MSA

```
human β     --------VHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQR    40
horse β     --------VQLSGEEKAAVLALWDKV--NEEEVGGEALGRLLVVYPWTQR    40
human α     ---------VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKT    41
horse α     ---------VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKT    41
lamprey g   PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQE    50
whale m     ---------VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLE    41
lupine l    --------GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKD    42
consensus             !*  **  * * ! **     * * *** * *   ! *
```

Conserved Histidine

```
human β     FFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDN-----LKGTF    85
horse β     FFDSFGDLSNPGAVMGNPKVKAHGKKVLHSFGEGVHHLDN-----LKGTF    85
human α     YFPHF-DLSH-----GSAQVKGHGKKVADALTNAVAHVDD-----MPNAL    80
horse α     YFPHF-DLSH-----GSAQVKAHGKKVGDALTLAVGHLDD-----LPGAL    80
lamprey g   FFPKFKGLTTADQLKKSADVRWHAERIINAVNDAVASMDDTEK--MSMKL    98
whale m     KFDRFKHLKTEAEMKASEDLKKHGVTVLTALGAILKKKGH-----HEAEL    86
lupine l    LFSFLKGTSEVPQ--NNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATL    90
consensus    !   * ***       **   ***!****   *    ** ***         *
```

Conserved Histidine

```
human β     ATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVA   135
horse β     AALSELHCDKLHVDPENFRLLGNVLVVVLARHFGKDFTPELQASYQKVVA   135
human α     SALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLA   130
horse α     SNLSDLHAHKLRVDPVNFKLLSHCLLSTLAVHLPNDFTPAVHASLDKFLS   130
lamprey g   RDLSGKHAKSFQVDPQYFKVLAAVIADTV---------AAGDAGFEKLMS   139
whale m     KPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALE   136
lupine l    KNLGSVHVSK-GVADAHFPVVKEAILKTIKEVVGAKWSEELNSAWTIAYD   139
consensus     !* *!* ** *** ** **    *   *** *     ***    *    *
```

```
human β     GVANALAHKYH------   146
horse β     GVANALAHKYH------   146
human α     SVSTVLTSKYR------   141
horse α     SVSTVLTSKYR------   141
lamprey g   MICILLRSAY-------   149
whale m     LFRKDIAAKYKELGYQG   153
lupine l    ELAIVIKKEMNDAA---   153
consensus    *    *   **
```

Progressive MSA (clustal)

```
human β     --------VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQR    40
horse β     --------VQLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQR    40
human α     ---------VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKT    41
horse α     ---------VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKT    41
lamprey g   PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQE    50
whale m     ---------VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLE    41
lupine l    --------GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKD    42
consensus             !*  **  * * ! **     * * *** * *   ! *

                                  Conserved Histidine
                                          ↓
human β     FFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDN-----LKGTF    85
horse β     FFDSFGDLSNPGAVMGNPKVKAHGKKVLHSFGEGVHHLDN-----LKGTF    85
human α     YFPHF-DLS-----HGSAQVKGHGKKVADALTNAVAHVDD-----MPNAL    80
horse α     YFPHF-DLS-----HGSAQVKAHGKKVGDALTLAVGHLDD-----LPGAL    80
lamprey g   FFPKFKGLTTADQLKKSADVRWHAERIINAVNDAVASMDDT--EKMSMKL    98
whale m     KFDRFKHLKTEAEMKASEDLKKHGVTVLTALGAILKKKGH-----HEAEL    86
lupine l    LFSFLKGTSEVP--QNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATL    90
consensus    !  * ***      **  ***!****  *   ** ***         *

            Conserved Histidine
                    ↓
human β     ATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVA   135
horse β     AALSELHCDKLHVDPENFRLLGNVLVVVLARHFGKDFTPELQASYQKVVA   135
human α     SALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLA   130
horse α     SNLSDLHAHKLRVDPVNFKLLSHCLLSTLAVHLPNDFTPAVHASLDKFLS   130
lamprey g   RDLSGKHAKSFQVDPQYFKVLAAVIADTVAAG---------DAGFEKLMS   139
whale m     KPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALE   136
lupine l    KNLGSVHVSKG-VADAHFPVVKEAILKTIKEVVGAKWSEELNSAWTIAYD   139
consensus     !* *!* ** *** ** **    *  *** *     ***    *   *

human β     GVANALAHKYH------   146
horse β     GVANALAHKYH------   146
human α     SVSTVLTSKYR------   141
horse α     SVSTVLTSKYR------   141
lamprey g   MICILLRSAY-------   149
whale m     LFRKDIAAKYKELGYQG   153
lupine l    ELAIVIKKEMNDAA---   153
consensus    *   *   **
```

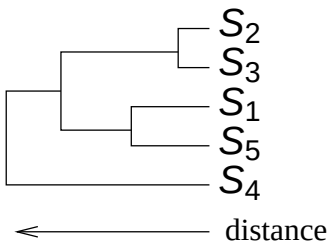
Progressive Multiple Sequence Alignment

Pairwise Phase

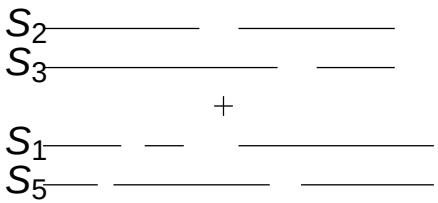
$\longrightarrow$

	S_1	S_2	S_3	S_4	S_5
S_1	0				
S_2	17	0			
S_3	17	4	0		
S_4	25	25	25	0	
S_5	10	17	17	25	0

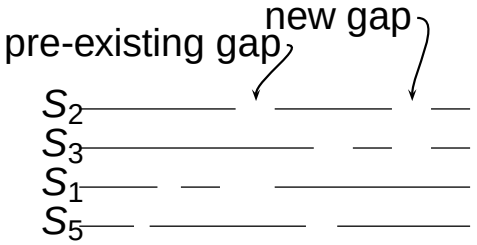
$\longrightarrow$



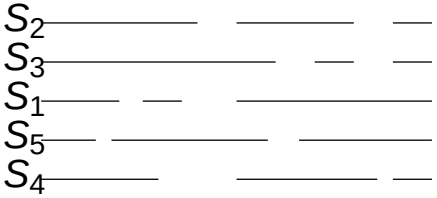
Multiple Sequence Phase



$\longrightarrow$



$\longrightarrow$



Exercise: Pairwise Alignment of Alignments

Alignments to be aligned:

1. AACGT
   A-CGT
2. AAGT
   A-GT

Score: match $\equiv$ +1; mismatch $\equiv$ -1; gap $\equiv$ -1

Solution

	- -	A A	A -	C C	G G	T T
- -	**0**	$\leftarrow -3$	$\leftarrow -6$	$\leftarrow -9$	$\leftarrow -12$	$\leftarrow -15$
A A	$\uparrow -3$	$\nwarrow$ **6**	$\leftarrow 3$	$\leftarrow 0$	$\leftarrow -3$	$\leftarrow -6$
A -	$\uparrow -6$	$\uparrow 3$	$\nwarrow$ **3**	$\leftarrow$ **0**	$\leftarrow -3$	$\leftarrow -6$
G G	$\uparrow -9$	$\uparrow 0$	$\uparrow 0$	$\nwarrow 1$	$\nwarrow$ **6**	$\leftarrow 3$
T T	$\uparrow -12$	$\uparrow -3$	$\uparrow -3$	$\uparrow\nwarrow -2$	$\uparrow 3$	$\nwarrow$ **12**

```
AACGT
A-CGT
AA-GT
A- -GT
```