

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

Bernhard Haubold & Thomas Wiehe

Outline

- 1 Why Multiple Sequence Alignment?**
- 2 Optimal Multiple Sequence Alignment**
- 3 Progressive Multiple Sequence Alignment**

Multiple Sequence Alignment

human β -----VHITPEEKSAVTAIWKYKN--VDEVGGEALGRLLVVYDWSQR 40
 horse β -----VQSGEKKAAVLRNDKVN--EEEVGGEALGRLLVVYDWSQR 40
 human α -----VSPADTKNWKRAWKYGAHAGEYGAEALEERMELSLFETPKT 41
 horse α -----VSAADTKNWKRAWKYGGHAGEYGAEALEERMELSLFETPKT 41
 lamprey g PIVDTGSVAPSSAAYTKIRSAAPLYSTETSETSDVDILVKFESTSTRAOE 50
 whale m -----VSEGEWQVLHLVWAKVERADVAGHGQDILIRLPLKSHPETLE 41
 lupine l -----GATEESQAALYKSSWEFPNANIPKHTHRFFILVLEIAAKD 42
 consensus ! * * * ! * * * * * * ! *

Conserved Histidine

↓
 human β FESPCGLIDTPDAVMGNPKVKAAGKKVLGFSDGLAHIDN-----LKGTf 85
 horse β FDSFGQLSNSGAVMGNPKVKAAGKKVLHSFGEDVHIDN-----LKGTf 85
 human α MYPHF DLS-----HGSAQEVGIGKRVADLTNAWAVDID-----MPNAE 80
 horse α YPHF DLS-----HGSAQEVKAAGKKVGDALTLAVGHIDD-----LPGAB 80
 lamprey g FDPFKGELTTADQLKKSADYRVAERINAVNDVASHEDDT-EKNSMK 98
 whale m KDRFKHLKTEAEMKASEDLKKRGVTALGAIKKGH-----HEAEL 86
 lupine l LSFALKGTSVEP-QNNPELQHAGKVFKLVYEAIQDQVTGVVVVTDAT 90
 consensus ! * * * * * * * * * * * * * * * * * *

Conserved Histidine

↓
 human β ATLSELELICDKLHVDFENEPRLPFGNVLVCVLAHHFGKEFTFPVCAAYQKVVA 135
 horse β AALSELELICDKLHVDFENEPRLPFGNVLVCVLAHHFGKEFTFPVCAAYQKVVA 135
 human α SALSDLHAAHKLIRVDEVNEKLLSHCLLVTLAHLFAEFTPAVHASLDKFLA 130
 horse α SNLSDLHAAHKLIRVDEVNEKLLSHCLLSTLAVHLPLNDFTPAVHASLDKFLS 130
 lamprey g RDLSGKHAKSFQVDBOYEVLAIAVIADTVHAG-----DAGFEKLM 139
 whale m KPLAQSHTAHKIPTKYLEFISEAIIHVVLHSRHPGEGADAQGMNKALE 136
 lupine l KNGSVYVSEK-VADAHFPVKEAILMTIKEVVGAKWSEELNSAWTIAYD 139
 consensus ! * * * ! * * * * * * * * * * * * * *

human β GIANAAAHKYH----- 146
 horse β GIANAAAHKYH----- 146
 human α SUSTVITSKVR----- 141
 horse α SUSTVITSKVR----- 141
 lamprey g MICKLERSAV----- 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 $\equiv 0$

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

Path in Hyperlattice

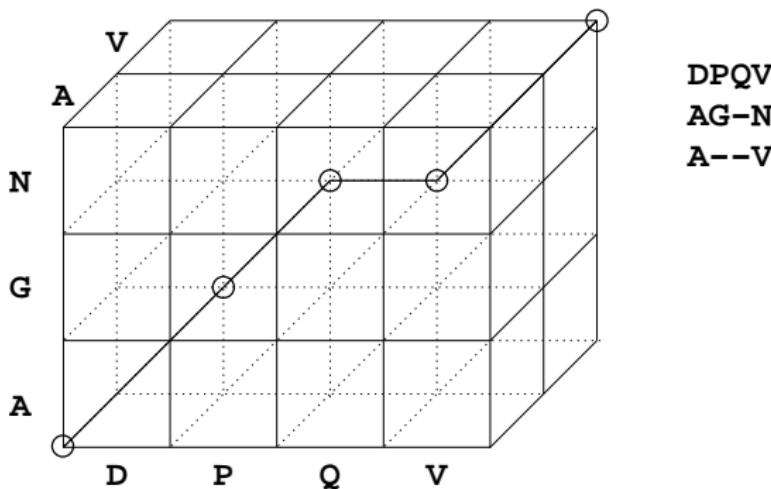
$(0, 0, 0, 0, 0, 0, 0) \rightarrow (0, 0, 0, 0, 1, 0, 0) \rightarrow \dots \rightarrow (0, 0, 0, 0, 8, 0, 0)$
 $\rightarrow (1, 1, 0, 0, 9, 0, 1) \rightarrow (2, 2, 1, 1, 10, 1, 2) \rightarrow \dots \rightarrow$
 $(147, 147, 141, 141, 149, 153, 153)$

Number of Vertices

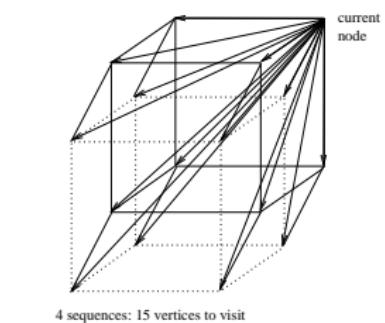
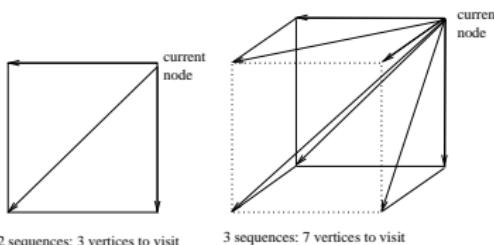
$$\prod_{i=1}^n (|\mathcal{S}_i| + 1)$$

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

Multiple Sequence Alignment by Dynamic Programming



Nodes in Alignment Hyperlattice



4 sequences: 15 vertices to visit

Optimal MSA

human β -----VHITPEEKSAVTAWKGKV-----NVDEVGGEALGRLLVVYIWNQR 40
 horse β -----VQSGEKKAAVLRNDKVV---NNEEVVGGEALGRLLVVYIWNQR 40
 human α -----VSPADTKTHWKRAKGKVGAHAGEYGAEAERERLPLSFTPKT 41
 horse α -----VSSAADTKTHWKRAKGKVGAHAGEYGAEAERERLPLSFTPKT 41
 lamprey g PIVDTGSVAPSSAAYTKIRSLPAPYSTVETSCVVDLIVKFESTSTRAOE 50
 whale m -----VSEGEWQVLHLVWAKVERADVAGHGQDILIRLFLPKSHPETLE 41
 lupine l -----GAATESQAAVVKSSWEEPNANIPKHTHRFFILVLEIAAKD 42
 consensus ! * * * ! * * * * * * ! *

Conserved Histidine

↓
 human β FESPFQDLSTPDAMVCNPKVKAHGKVLGCFSQDGLAHLDN-----LKGTF 85
 horse β FDSFGQLSNSGAVMGNPKVKAHGKVLHSFGEDVHHDN-----LKGTF 85
 human α YPHFDLSH-----GSAQVEGIGKRVADLTNAWAVD-----MPNAE 80
 horse α YPHFDLSH-----GSAQVEGIGKRVADLTNAWAVD-----MPNAE 80
 lamprey g FPKFKGLTTADQLKKSADLVRV-----GSAQVEGIGKRVADLTNAWAVD 98
 whale m KDRFKHLKTEAEMLKASEDLKKRGTVTLALGAIKKKGH-----REAEED 86
 lupine l LSFALKGTSEVFPQ--NNPELQAHAGKVFKLVYEAIQDQVTGVVVVTDAT 90
 consensus ! *

Conserved Histidine

↓
 human β ATLSEELCDKLVHDFENEPRLPGNVLVCVLAHHFGKEFTPPVCAAYQKVVA 135
 horse β AALSEELCDKLVHDFENEPRLPGNVLVCVLAHHFGKEFTPPVCAAYQKVVA 135
 human α SALSDLHAAHKLIRVDEVNPKLLSHCLLVTLAHLFAEFTPAVHASLDKFLA 130
 horse α SNLSDLHAAHKLIRVDEVNPKLLSHCLLSTLAVLPNDFTPAVHASLDKFLA 130
 lamprey g RDLSGKIAKSFQVDBQYEVKLAIAVIAADTV-----AAGDAGFEKLM 139
 whale m KPLAQSHATRKHIPTKYLEFISEAIIHVVLHSRHPGEGADAQGMNKALE 136
 lupine l KNGSVYHVSAGVADAHGPVKEAILMKTIKEVVGAKWSEELNSAWTIAYD 139
 consensus ! * * * ! *

human β GYANAIAHKYH----- 146
 horse β GWANAIAHKYH----- 146
 human α SUSTVITSKVR----- 141
 horse α SUSTVITSKVR----- 141
 lamprey g MICILERSAV----- 149
 whale m LFRKDIAAKYELGYOG 153
 lupine l ELAIVIKKEMNDAA--- 153
 consensus * * *

Progressive MSA (clustal)

human β -----VHITPEEKSAVTAIWKYKN--VDEVGGEALGRLLVVVFWQR 40
 horse β -----VQSGEKKAAVLRNDKVN--EEEVGGEALGRLLVVVFWQR 40
 human α -----VSPADKTHWKRNGKYGAHAGEYGAEALEERMELSLFETKT 41
 horse α -----VSAADKTHWKRNSKGHGAEYGAEALEERMELSLFETKT 41
 lamprey g PIVDTGSVAPSSAAYTKIRSAAPLYSTVETSETSDVILVKFETSTRAOE 50
 whale m -----VSEGEWQLVLHVWAKVERADVAGHQDILIRLFLPKSHPETLE 41
 lupine l -----GATEESQAALVKSSWEEPNNANIPKHTHRFFILVLEIAAKD 42
 consensus ! * * * ! * * * * * * ! *

Conserved Histidine

↓
 human β FESPGCGLSTPPDAVMGNPKVKAAGKKVLGFSDGLAHIDN-----LKGTF 85
 horse β FDSFGGLSNSGAVMGNPKVKAAGKKVLHSFGEDVHIDN-----LKGTF 85
 human α YPHFDLS----HGSAQEVGIGKRVADLTNAWAVD-----MPNAE 80
 horse α YPHFDLS----HGSAQEVKAAGKKVGDALTLNAWAVD-----LPGAB 80
 lamprey g DRFKHLKTEAEMKASEDLKKRGVTALGAIKKGH-----HEAEL 98
 whale m DRFKHLKTEAEMKASEDLKKRGVTALGAIKKGH-----HEAEL 86
 lupine l DRFKHLKTSVEP-QNNPELQAHAGKVFKLVYEAIQDQVTGVVVTDAT 90
 consensus ! *

Conserved Histidine

↓
 human β ATLSELELICDKLHVDFENEPRLPFGNVLVCVLAHHFGKEFTFPVCAAYQKVVA 135
 horse β AALSELELICDKLHVDFENEPRLPFGNVLVCVLAHHFGKEFTFPVCAAYQKVVA 135
 human α SALSDLHAAHKLIRVDEVNPKLLSHCLLVTLAHLFAEFTPAVHASLDKFLA 130
 horse α SNLSDLHAAHKLIRVDEVNPKLLSHCLLSTLAVHLPNDFTPAVHASLDKFLS 130
 lamprey g RDLSGKHAKSFQVDBQEYKVLAAVIADTVHAG-----DAGFEKLM 139
 whale m KPLAQSHTAHKIPTKYLEFISEAIIHVVLHSRHPGEGADAQGMNKALE 136
 lupine l KNGSVHVSKEG-VADAHFPVKEAILMTIKEVVGAKWSEELNSAWTIAYD 139
 consensus ! * * * ! * * * * * * * * * * * * * * * * *

human β GIANAIAHKYH----- 146
 horse β GIANAIAHKYH----- 146
 human α SUSTVITSKVR----- 141
 horse α SUSTVITSKVR----- 141
 lamprey g MICILERSAV----- 149
 whale m LFRKDIAAKYKELGYQG 153
 lupine l ELAIVIKKEMNDAA--- 153
 consensus * * *



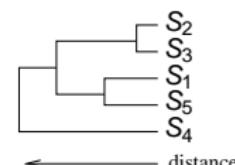
Progressive Multiple Sequence Alignment

S_1
 S_2
 S_3
 S_4
 S_5



Pairwise Phase

	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



S_2
 S_3
+
 S_1
 S_5



Multiple Sequence Phase

pre-existing gap \nearrow new gap \searrow
 S_2
 S_3
 S_1
 S_5



S_2
 S_3
 S_1
 S_5
 S_4

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	C	G	T
-	-	A	-	C	G	T
-	0	$\leftarrow -3$	$\leftarrow -6$	$\leftarrow -9$	$\leftarrow -12$	$\leftarrow -15$
A	$\uparrow -3$	$\nwarrow \text{ } \mathbf{6}$	$\leftarrow 3$	$\leftarrow 0$	$\leftarrow -3$	$\leftarrow -6$
A	$\uparrow -6$	$\uparrow 3$	$\nwarrow \mathbf{3}$	$\leftarrow \mathbf{0}$	$\leftarrow -3$	$\leftarrow -6$
A	$\uparrow -9$	$\uparrow 0$	$\uparrow 0$	$\nwarrow 1$	$\nwarrow \mathbf{6}$	$\leftarrow 3$
G	$\uparrow -12$	$\uparrow -3$	$\uparrow -3$	$\nwarrow \nwarrow -2$	$\uparrow 3$	$\nwarrow \mathbf{12}$
G						
T						
T						

AACGT
 A-CGT
 AA-GT
 A--GT