Clustal 1988-2024

Des Higgins

UCD

Multiple Alignment?

- Align 3 or more sequences together

```
TPIS_THEPX      TAKDANEVIKAIRNTIASLYGKEKADLVRIQYGGSVKPENISELMAESDIDGALVGGASL
TPIS_GEOTN      TAEDANNVCGHIRSVVARLFGAEAAEAIRIQYGGSVKPENISDFLAQEHIDGALVGGASL
TPIS_SYNP2      -ATEANRVIGLIR---EKLTNKN----VTIQYGGSVKPNNVDEIMAQPEIDGALVGGASL
                 * :**.*   **     * . :    : *********:*:.:::*: .***********
```

Homologous residues lined up in columns

Needed because of

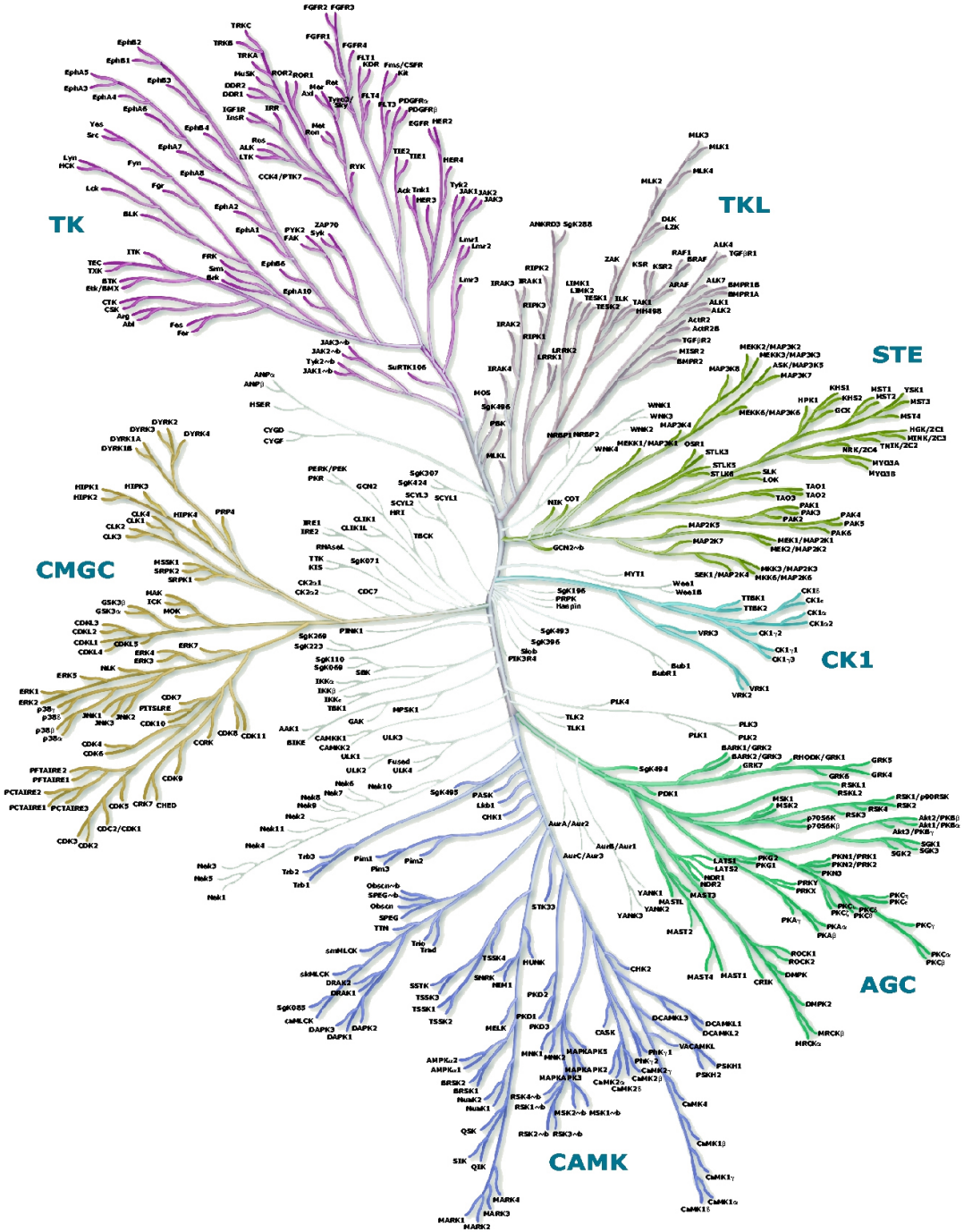
Orthologues from different species

But mainly:

Paralogues from Gene duplications

Multi-gene families

e.g. humans have approx. 500 protein kinases



Human Protein Kinases

The human kinome comprises 40 atypical PKs and 478 classical PKs. The latter consist of 388 serine/threonine kinases, 90 tyrosine kinases and 50 sequences which lack a functional catalytic site.

(Manning et al., Science, 2002)

Globin Multiple Alignment

```
Human beta          --------VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLST
Horse beta          --------VQLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDLSN
Human alpha         ---------VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-
Horse alpha         ---------VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF-DLS-
Whale myoglobin     ---------VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKT
Lamprey globin      PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLTT
Lupin globin        --------GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE
                              *:  :   :   *  .           :  .:   * :   *  :   .

Human beta          PDAVMGNPKVKAHGKKVLGAFSDGLAHLDN-----LKGTFATLSELHCDKLHVDPENFRL
Horse beta          PGAVMGNPKVKAHGKKVLHSFGEGVHHLDN-----LKGTFAALSELHCDKLHVDPENFRL
Human alpha         ----HGSAQVKGHGKKVADALTNAVAHVDD-----MPNALSALSDLHAHKLRVDPVNFKL
Horse alpha         ----HGSAQVKAHGKKVGDALTLAVGHLDD-----LPGALSNLSDLHAHKLRVDPVNFKL
Whale myoglobin     EAEMKASEDLKKHGVTVLTALGAILKKKGH-----HEAELKPLAQSHATKHKIPIKYLEF
Lamprey globin      ADQLKKSADVRWHAERIINAVNDAVASMDDT--EKMSMKLRDLSGKHAKSFQVDPQYFKV
Lupin globin        VP--QNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKGVAD-AHFPV
                          . .:: *.  :   .                    :  *.  *  .       : .

Human beta          LGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------
Horse beta          LGNVLVVVLARHFGKDFTPELQASYQKVVAGVANALAHKYH------
Human alpha         LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
Horse alpha         LSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTSKYR------
Whale myoglobin     ISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
Lamprey globin      LAAVIADTVAAG---D------AGFEKLMSMICILLRSAY-------
Lupin globin        VKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
```

Globin Multiple Alignment

```
Human beta          --------VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLST
Horse beta          --------VQLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDLSN
Human alpha         ---------VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-
Horse alpha         ---------VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF-DLS-
Whale myoglobin     ---------VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKT
Lamprey globin      PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLTT
Lupin globin        --------GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE
                              *:  :   :   *  .           :  .:   * :   *  :   .

Human beta          PDAVMGNPKVKAHGKKVLGAFSDGLAHLDN-----LKGTFATLSELHCDKLHVDPENFRL
Horse beta          PGAVMGNPKVKAHGKKVLHSFGEGVHHLDN-----LKGTFAALSELHCDKLHVDPENFRL
Human alpha         ----HGSAQVKGHGKKVADALTNAVAHVDD-----MPNALSALSDLHAHKLRVDPVNFKL
Horse alpha         ----HGSAQVKAHGKKVGDALTLAVGHLDD-----LPGALSNLSDLHAHKLRVDPVNFKL
Whale myoglobin     EAEMKASEDLKKHGVTVLTALGAILKKKGH-----HEAELKPLAQSHATKHKIPIKYLEF
Lamprey globin      ADQLKKSADVRWHAERIINAVNDAVASMDDT--EKMSMKLRDLSGKHAKSFQVDPQYFKV
Lupin globin        VP--QNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKGVAD-AHFPV
                          . .:: *.  :   .                   :  *.  *  .      : .

Human beta          LGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------
Horse beta          LGNVLVVVLARHFGKDFTPELQASYQKVVAGVANALAHKYH------
Human alpha         LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
Horse alpha         LSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTSKYR------
Whale myoglobin     ISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
Lamprey globin      LAAVIADTVAAG---D------AGFEKLMSMICILLRSAY-------
Lupin globin        VKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
```

Globin Multiple Alignment

```
Human beta          --------VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLST
Horse beta          --------VQLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDLSN
Human alpha         ---------VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-
Horse alpha         ---------VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF-DLS-
Whale myoglobin     ---------VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKT
Lamprey globin      PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLTT
Lupin globin        --------GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE
                              *:  :   :   *  .           :  .:   * :   *  :   .

Human beta          PDAVMGNPKVKAHGKKVLGAFSDGLAHLDN-----LKGTFATLSELHCDKLHVDPENFRL
Horse beta          PGAVMGNPKVKAHGKKVLHSFGEGVHHLDN-----LKGTFAALSELHCDKLHVDPENFRL
Human alpha         ----HGSAQVKGHGKKVADALTNAVAHVDD-----MPNALSALSDLHAHKLRVDPVNFKL
Horse alpha         ----HGSAQVKAHGKKVGDALTLAVGHLDD-----LPGALSNLSDLHAHKLRVDPVNFKL
Whale myoglobin     EAEMKASEDLKKHGVTVLTALGAILKKKGH-----HEAELKPLAQSHATKHKIPIKYLEF
Lamprey globin      ADQLKKSADVRWHAERIINAVNDAVASMDDT--EKMSMKLRDLSGKHAKSFQVDPQYFKV
Lupin globin        VP--QNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKGVAD-AHFPV
                          . .:: *.  :   .                    :  *.  *  .       : .

Human beta          LGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------
Horse beta          LGNVLVVVLARHFGKDFTPELQASYQKVVAGVANALAHKYH------
Human alpha         LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
Horse alpha         LSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTSKYR------
Whale myoglobin     ISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
Lamprey globin      LAAVIADTVAAG---D------AGFEKLMSMICILLRSAY-------
Lupin globin        VKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
```

Alpha helices

Globin Multiple Alignment

```
Human beta          --------VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLST
Horse beta          --------VQLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDLSN
Human alpha         ---------VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-
Horse alpha         ---------VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF-DLS-
Whale myoglobin     ---------VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKT
Lamprey globin      PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLTT
Lupin globin        --------GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE
                              *:  :   :   *  .           :  .:   * :   *  :   .

Human beta          PDAVMGNPKVKAHGKKVLGAFSDGLAHLDN-----LKGTFATLSELHCDKLHVDPENFRL
Horse beta          PGAVMGNPKVKAHGKKVLHSFGEGVHHLDN-----LKGTFAALSELHCDKLHVDPENFRL
Human alpha         ----HGSAQVKGHGKKVADALTNAVAHVDD-----MPNALSALSDLHAHKLRVDPVNFKL
Horse alpha         ----HGSAQVKAHGKKVGDALTLAVGHLDD-----LPGALSNLSDLHAHKLRVDPVNFKL
Whale myoglobin     EAEMKASEDLKKHGVTVLTALGAILKKKGH-----HEAELKPLAQSHATKHKIPIKYLEF
Lamprey globin      ADQLKKSADVRWHAERIINAVNDAVASMDDT--EKMSMKLRDLSGKHAKSFQVDPQYFKV
Lupin globin        VP--QNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKGVAD-AHFPV
                          . .:: *.  :   .                    :  *.  *  .       : .

Human beta          LGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------
Horse beta          LGNVLVVVLARHFGKDFTPELQASYQKVVAGVANALAHKYH------
Human alpha         LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
Horse alpha         LSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTSKYR------
Whale myoglobin     ISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
Lamprey globin      LAAVIADTVAAG---D------AGFEKLMSMICILLRSAY-------
Lupin globin        VKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
```

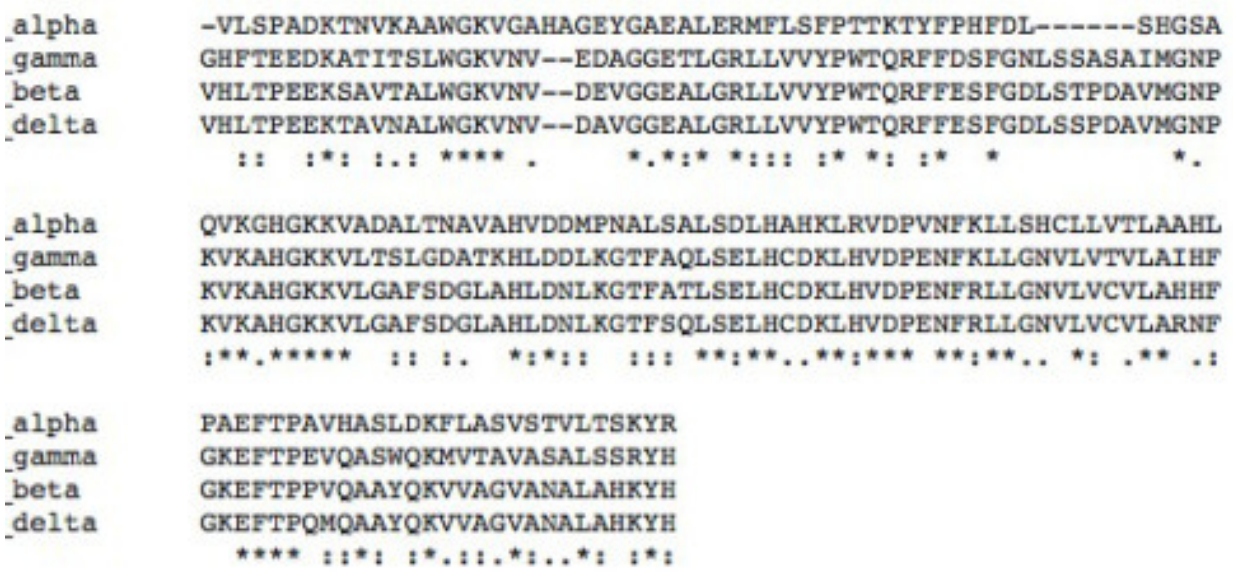
Haem binding Histidines

Globin Multiple Alignment

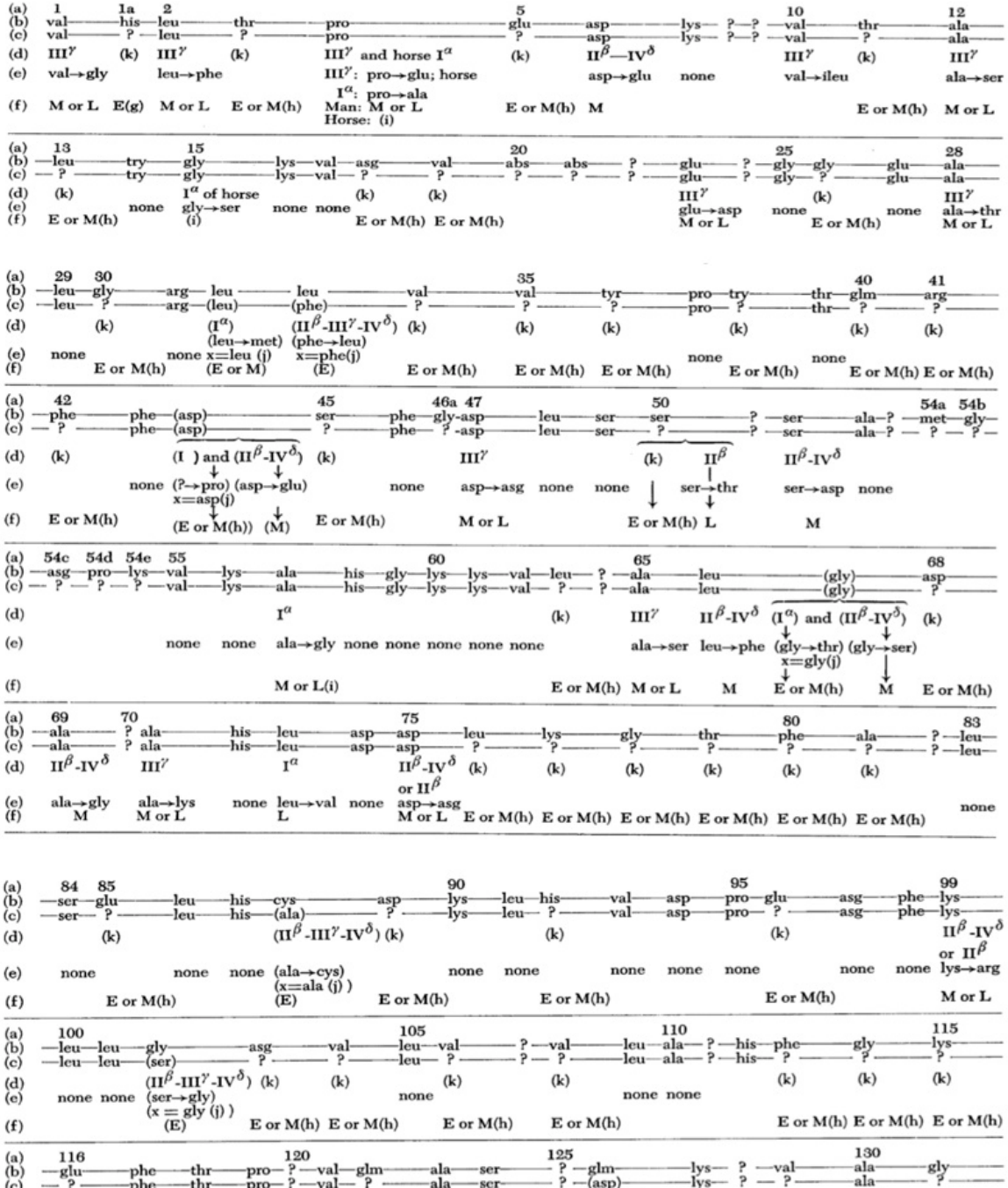
```
Human beta          --------VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLST
Horse beta          --------VQLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDLSN
Human alpha         ---------VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-
Horse alpha         ---------VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF-DLS-
Whale myoglobin     ---------VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKT
Lamprey globin      PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLTT
Lupin globin        --------GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE
                              *:  :   :   *  .           :  .:   * :   *  :   .

Human beta          PDAVMGNPKVKAHGKKVLGAFSDGLAHLDN-----LKGTFATLSELHCDKLHVDPENFRL
Horse beta          PGAVMGNPKVKAHGKKVLHSFGEGVHHLDN-----LKGTFAALSELHCDKLHVDPENFRL
Human alpha         ----HGSAQVKGHGKKVADALTNAVAHVDD-----MPNALSALSDLHAHKLRVDPVNFKL
Horse alpha         ----HGSAQVKAHGKKVGDALTLAVGHLDD-----LPGALSNLSDLHAHKLRVDPVNFKL
Whale myoglobin     EAEMKASEDLKKHGVTVLTALGAILKKKGH-----HEAELKPLAQSHATKHKIPIKYLEF
Lamprey globin      ADQLKKSADVRWHAERIINAVNDAVASMDDT--EKMSMKLRDLSGKHAKSFQVDPQYFKV
Lupin globin        VP--QNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKGVAD-AHFPV
                          . .:: *.  :   .                    :  *.  *  .       : .

Human beta          LGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------
Horse beta          LGNVLVVVLARHFGKDFTPELQASYQKVVAGVANALAHKYH------
Human alpha         LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
Horse alpha         LSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTSKYR------
Whale myoglobin     ISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
Lamprey globin      LAAVIADTVAAG---D------AGFEKLMSMICILLRSAY-------
Lupin globin        VKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
```

The first MSA?

- Thanks to Lior Pachter
  https://liorpachter.wordpress.com/2013/12/08/molecular-restoration-studies-of-extinct-forms-of-life/
- L. Pauling and E. Zuckerkandl (1963)
  Chemical Paleogenetics: Molecular ‘restoration studies’ of extinct forms of life. Acta Chem. Scand. 17, S9–S16.

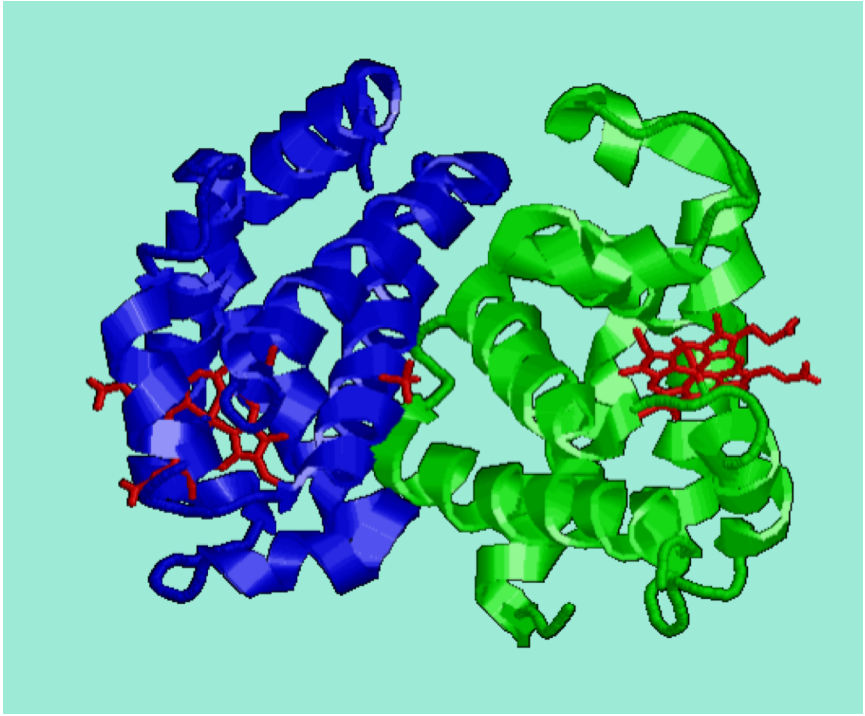


Pauling and Zuckerkandl, Japan, 1986


(a) 1 1a 2 5 10 12
(b) val his leu thr pro glu asp lys ? ? val thr ala
(c) val ? leu ? pro ? asp lys ? ? val ? ala
(d) III$^{\gamma}$ (k) III$^{\gamma}$ (k) III$^{\gamma}$ and horse I$^{\alpha}$ (k) II$^{\beta}$—IV$^{\delta}$ III$^{\gamma}$ (k) III$^{\gamma}$
(e) val→gly leu→phe III$^{\gamma}$: pro→glu; horse I$^{\alpha}$: pro→ala asp→glu none val→ileu ala→ser
(f) M or L E(g) M or L E or M(h) Man: M or L Horse: (i) E or M(h) M E or M(h) M or L
(a) 13 15 20 25 28
(b) leu try gly lys val asg val abs abs ? glu ? gly gly glu ala
(c) ? try gly lys val ? ? ? ? glu ? gly ? glu ala
(d) (k) I$^{\alpha}$ of horse (k) (k) III$^{\gamma}$ (k) III$^{\gamma}$
(e) none gly→ser none none glu→asp none none ala→thr
(f) E or M(h) (i) E or M(h) E or M(h) M or L E or M(h) M or L
(a) 29 30 35 40 41
(b) leu gly arg leu leu val val tyr pro try thr glu arg
(c) leu ? arg (leu) (phe) ? ? ? pro ? thr ? ? ?
(d) (k) (I$^{\alpha}$) (leu→met) (II$^{\beta}$-III$^{\gamma}$-IV$^{\delta}$) (phe→leu) (k) (k) (k) (k) (k) (k)
(e) none none x=leu (j) x=phe(j) none none
(f) E or M(h) (E or M) (E) E or M(h) E or M(h) E or M(h) E or M(h) E or M(h) E or M(h)
(a) 42 45 46a 47 50 54a 54b
(b) phe phe (asp) ser phe gly-asp leu ser ser ? ser ala ? met gly
(c) ? phe (asp) ? phe ? -asp leu ser ? ? ser ala ? ? ?
(d) (k) (I) and (II$^{\beta}$-IV$^{\delta}$) (k) III$^{\gamma}$ (k) II$^{\beta}$ II$^{\beta}$-IV$^{\delta}$
(e) none (?→pro) (asp→glu) x=asp(j) none asp→asg none none ser→thr ser→asp none
(f) E or M(h) (E or M(h)) (M) E or M(h) M or L E or M(h) L M
(a) 54c 54d 54e 55 60 65 68
(b) asg pro lys val lys ala his gly lys lys val leu ? ala leu (gly) asp
(c) ? ? ? val lys ala his gly lys lys val ? ? ala leu (gly) ?
(d) I$^{\alpha}$ (k) III$^{\gamma}$ II$^{\beta}$-IV$^{\delta}$ (I$^{\alpha}$) and (II$^{\beta}$-IV$^{\delta}$) (k)
(e) none none ala→gly none none none none none ala→ser leu→phe (gly→thr) (gly→ser) x=gly(j)
(f) M or L(i) E or M(h) M or L M E or M(h) M E or M(h)
(a) 69 70 75 80 83
(b) ala ? ala his leu asp asp leu lys gly thr phe ala ? leu
(c) ala ? ala his leu asp asp ? ? ? ? ? ? ? leu
(d) II$^{\beta}$-IV$^{\delta}$ III$^{\gamma}$ I$^{\alpha}$ II$^{\beta}$-IV$^{\delta}$ or II$^{\beta}$ (k) (k) (k) (k) (k) (k)
(e) ala→gly ala→lys none leu→val none asp→asg none
(f) M M or L L M or L E or M(h) E or M(h) E or M(h) E or M(h) E or M(h) E or M(h)
(a) 84 85 90 95 99
(b) ser glu leu his cys asp lys leu his val asp pro glu asg phe lys
(c) ser ? leu his (ala) ? lys leu ? val asp pro ? asg phe lys
(d) (k) (II$^{\beta}$-III$^{\gamma}$-IV$^{\delta}$) (k) (k) (k) II$^{\beta}$-IV$^{\delta}$ or II$^{\beta}$
(e) none none none (ala→cys) (x=ala (j)) none none none none none none none lys→arg
(f) E or M(h) (E) E or M(h) E or M(h) E or M(h) M or L
(a) 100 105 110 115
(b) leu leu gly asg val leu val ? val leu ala ? his phe gly lys
(c) leu leu (ser) ? ? leu ? ? ? ? leu ala ? his ? ? ? ?
(d) (II$^{\beta}$-III$^{\gamma}$-IV$^{\delta}$) (k) (k) (k) (k) (k) (k) (k)
(e) none none (ser→gly) (x = gly (j)) none none none
(f) (E) E or M(h) E or M(h) E or M(h) E or M(h) E or M(h) E or M(h) E or M(h)
(a) 116 120 125 130
(b) glu phe thr pro ? val glm ala ser ? glm lys ? val ala gly
(c) ? phe thr pro ? val ? ala ser ? (asp) lys ? ? ala ?

```
VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNP
-VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-----HGSA
  * *  *  * * ****     * * *** *     * *   *  * ***      *

KVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHF
QVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHL
 ** *****  *     ** *        ** **  ** *** ** **   *   ** *

GKEFTPPVQAAYQKVVAGVANALAHKYH
PAEFTPAVHASLDKFLASVSTVLTSKYR
  **** * *   *  * *   *  **
```

- Dynamic Programming
  - Needleman and Wunsch, 1970
  - $O(L^2)$ algorithm
- Maximise score
  (or minimise distance)
  - Gap penalties
  - Amino acid weight matrix



Blosum 62 Matrix

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

- +ve: likely substitutions -ve: unlikely subs.
- from counting matches in blocks of alignment
- Henikoff and Henikoff, 1992

			A		T		T		A		C		C
	0	→	0.1	→	0.3	→	0.5	→	0.6	→	0.7	→	0.8
	↓	↘	↓		↓		↓	↘	↓		↓		↓
G	**0.1**	→	0.2	→	0.4		0.6	→	0.7	→	0.8	→	0.9
	↓	↘						↘					
A	0.3		**0.1**	→	0.3	→	0.5	→	0.6	→	0.7	→	0.8
	↓		↓	↘		↘							
T	0.4		0.2		**0.1**	→	0.3	→	0.4	→	0.5	→	0.6
	↓		↓		↓		↓	↘	↓		↓		↓
G	0.5		0.3		**0.2**	→	0.4	→	0.5	→	0.6	→	0.7
	↓		↓	↘	↓	↘							
T	0.6		0.4		0.3		**0.2**	→	0.3	→	0.4	→	0.5
	↓	↘			↓		↓	↘					
A	0.8		0.6		0.5		0.4		**0.2**	→	0.3	→	0.4
	↓	↘	↓		↓		↓	↘	↓	↘		↘	
A	1		0.8		0.7		0.6		0.4		**0.4**	→	0.5
	↓		↓		↓	↘			↓	↘		↘	
C	1.2		1		0.9		0.8		0.6		0.4		**0.4**
	↓	↘	↓		↓		↓	↘			↓	↘	↓
A	1.4		1.2		1.1		1		0.8		0.6		**0.6**

```
Human beta         --------VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLST
Horse beta         --------VQLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDLSN
Human alpha        ---------VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-
Horse alpha        ---------VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF-DLS-
Whale myoglobin    ---------VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKT
Lamprey globin     PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLTT
Lupin globin       --------GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE
                             *:  :   :   *  .           :  .:   * :   *  :   .

Human beta         PDAVMGNPKVKAHGKKVLGAFSDGLAHLDN-----LKGTFATLSELHCDKLHVDPENFRL
Horse beta         PGAVMGNPKVKAHGKKVLHSFGEGVHHLDN-----LKGTFAALSELHCDKLHVDPENFRL
Human alpha        ----HGSAQVKGHGKKVADALTNAVAHVDD-----MPNALSALSDLHAHKLRVDPVNFKL
Horse alpha        ----HGSAQVKAHGKKVGDALTLAVGHLDD-----LPGALSNLSDLHAHKLRVDPVNFKL
Whale myoglobin    EAEMKASEDLKKHGVTVLTALGAILKKKGH-----HEAELKPLAQSHATKHKIPIKYLEF
Lamprey globin     ADQLKKSADVRWHAERIINAVNDAVASMDDT--EKMSMKLRDLSGKHAKSFQVDPQYFKV
Lupin globin       VP--QNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKGVAD-AHFPV
                         . .:: *.  :   .                 :  *.  *  .      : .

Human beta         LGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------
Horse beta         LGNVLVVVLARHFGKDFTPELQASYQKVVAGVANALAHKYH------
Human alpha        LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
Horse alpha        LSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTSKYR------
Whale myoglobin    ISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
Lamprey globin     LAAVIADTVAAG---D------AGFEKLMSMICILLRSAY-------
Lupin globin       VKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
                   :   :  .:      .      ..       .   :
```

$$\sum_{i=2}^{N}\sum_{j=1}^{i-1} W_{ij} D_{ij}$$

Time $O(L^N)$

```
Human beta          --------VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLST
Horse beta          --------VQLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDLSN
Human alpha         ---------VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-
Horse alpha         ---------VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF-DLS-
Whale myoglobin     ---------VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKT
Lamprey globin      PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLTT
Lupin globin        --------GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE
                              *:  :   :   *  .           :  .:   * :   *  :   .

Human beta          PDAVMGNPKVKAHGKKVLGAFSDGLAHLDN-----LKGTFATLSELHCDKLHVDPENFRL
Horse beta          PGAVMGNPKVKAHGKKVLHSFGEGVHHLDN-----LKGTFAALSELHCDKLHVDPENFRL
Human alpha         ----HGSAQVKGHGKKVADALTNAVAHVDD-----MPNALSALSDLHAHKLRVDPVNFKL
Horse alpha         ----HGSAQVKAHGKKVGDALTLAVGHLDD-----LPGALSNLSDLHAHKLRVDPVNFKL
Whale myoglobin     EAEMKASEDLKKHGVTVLTALGAILKKKGH-----HEAELKPLAQSHATKHKIPIKYLEF
Lamprey globin      ADQLKKSADVRWHAERIINAVNDAVASMDDT--EKMSMKLRDLSGKHAKSFQVDPQYFKV
Lupin globin        VP--QNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKGVAD-AHFPV
                          . .:: *.  :   .                 :  *.  *  .       : .

Human beta          LGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------
Horse beta          LGNVLVVVLARHFGKDFTPELQASYQKVVAGVANALAHKYH------
Human alpha         LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
Horse alpha         LSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTSKYR------
Whale myoglobin     ISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
Lamprey globin      LAAVIADTVAAG---D------AGFEKLMSMICILLRSAY-------
Lupin globin        VKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
                    :   :  .:      .      ..       .   :
```

Sequences	Time
2	1 second
3	150 seconds
4	6.25 hours
5	39 days
6	16 years
7	2404 years

Time $O(L^N)$

The first automatic MSA?

- David Sankoff, 1973!

Evolution of 5S RNA and the Non-randomness of Base Replacement

SEQUENCES of 5S rRNA have been published for five widely divergent organisms: *Escherichia coli*, *Pseudomonas fluorescens*, yeast, human KB carcinoma, and *Xenopus laevis*[1–5]. The question arises whether sequences for the ancestors of these organisms, represented by X, Y, and Z in Fig. 1, can be reconstructed to any degree of confidence; were this possible, statistical analysis of mutation types would become feasible. Such reconstructions are well known in protein studies[6] and for tRNAs[7], but they are somewhat more difficult for 5S rRNA.

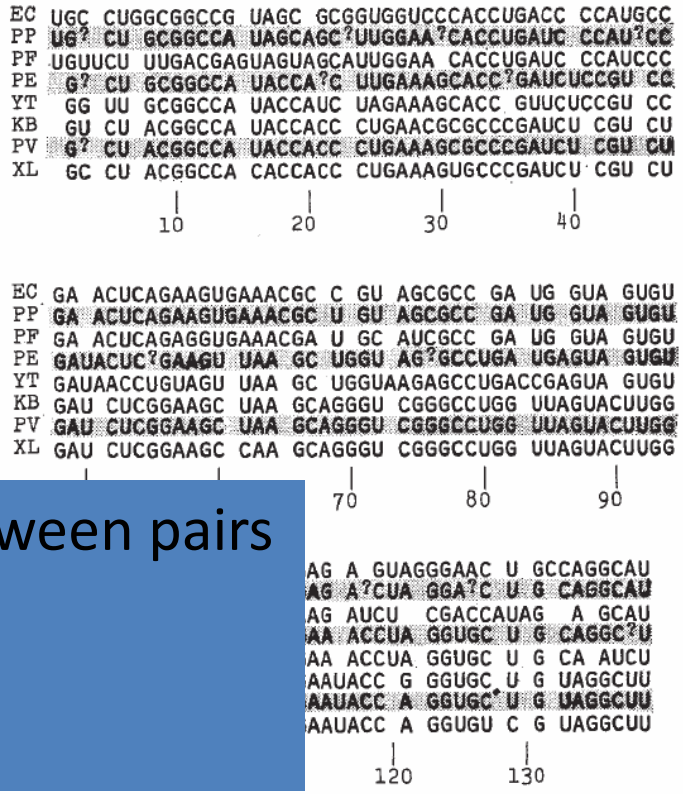


Fig. 2 Alignment of known and reconstructed 5S RNA sequences. From top to bottom: EC, *E. coli*; PP, proto-prokaryote; PF, *P. fluorescens*; PE, proto-eukaryote; YT, yeast; KB, human KB carcinoma; PV, proto-vertebrate; XL, *Xenopus*. Reconstructed sequences are shaded; ?, undetermined base or gap.

- Reconstruct ancestral seqs between pairs
  - Dynamic programming
- Align in steps

NATURE NEW

DAVID SANKOFF
CRISTIANE MOREL
ROBERT J. CEDERGREN

Centre de Recherches Mathématiques and
Département de Biochimie,
Université de Montréal,
Case Postale 6128, *Montréal* 101

Paulien Hogeweg

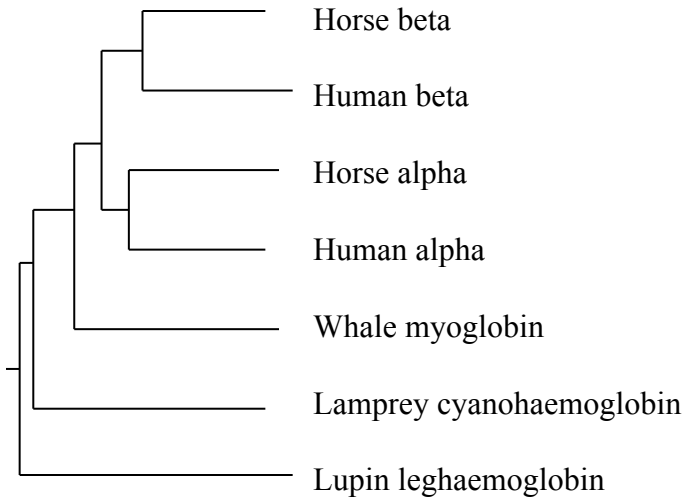
- Utrecht, The Netherlands
- “Bioinformatics”
  - Hesper B., Hogeweg P. (1970.) "Bioinformatica: een werkconcept", *Kameleon*, 1(6): 28–29.
- Automated MSA 1984
  - Hogeweg, P. and B. Hesper B. (**1984**) The alignment of sets of sequences and the construction of phyletic trees: an integrated method. J Mol Evol 20: 175-186.
  - Align sequences together gradually

```
Human beta          --------VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLST
Horse beta          --------VQLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDLSN
Human alpha         ---------VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-
Horse alpha         ---------VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF-DLS-
Whale myoglobin     ---------VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKT
Lamprey globin      PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLTT
Lupin globin        --------GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE
                              *:  :   :   *  .           :  .:   * :   *  :   .

Human beta          PDAVMGNPKVKAHGKKVLGAFSDGLAHLDN-----LKGTFATLSELHCDKLHVDPENFRL
Horse beta          PGAVMGNPKVKAHGKKVLHSFGEGVHHLDN-----LKGTFAALSELHCDKLHVDPENFRL
Human alpha         ----HGSAQVKGHGKKVADALTNAVAHVDD-----MPNALSALSDLHAHKLRVDPVNFKL
Horse alpha         ----HGSAQVKAHGKKVGDALTLAVGHLDD-----LPGALSNLSDLHAHKLRVDPVNFKL
Whale myoglobin     EAEMKASEDLKKHGVTVLTALGAILKKKGH-----HEAELKPLAQSHATKHKIPIKYLEF
Lamprey globin      ADQLKKSADVRWHAERIINAVNDAVASMDDT--EKMSMKLRDLSGKHAKSFQVDPQYFKV
Lupin globin        VP--QNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKGVAD-AHFPV
                          . .:: *.  :   .                 :  *.  *  .       : .

Human beta          LGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------
Horse beta          LGNVLVVVLARHFGKDFTPELQASYQKVVAGVANALAHKYH------
Human alpha         LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
Horse alpha         LSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTSKYR------
Whale myoglobin     ISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
Lamprey globin      LAAVIADTVAAG---D------AGFEKLMSMICILLRSAY-------
Lupin globin        VKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
                    :   :  .:      .      ..       .   :
```

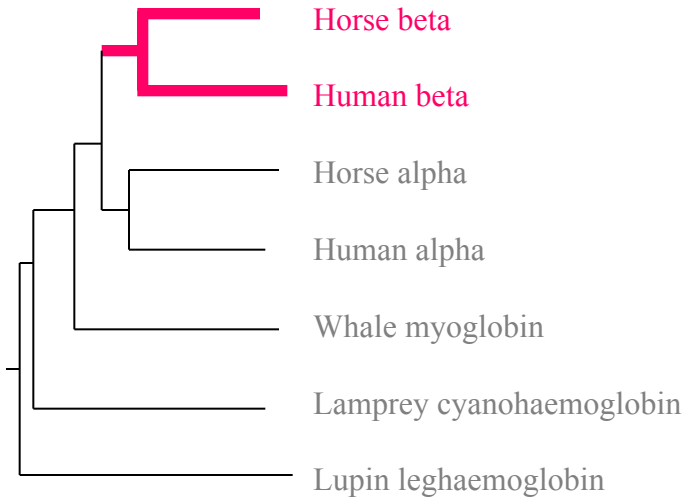
Progressive Alignment:

Barton and Sternberg, 1987
Florence Corpet, 1988
Feng and Doolittle, 1987
Jotun Hein, 1989
Higgins and Sharp, 1988, 1989
Hogeweg and Hesper, 1984
Willie Taylor, 1987, 1988

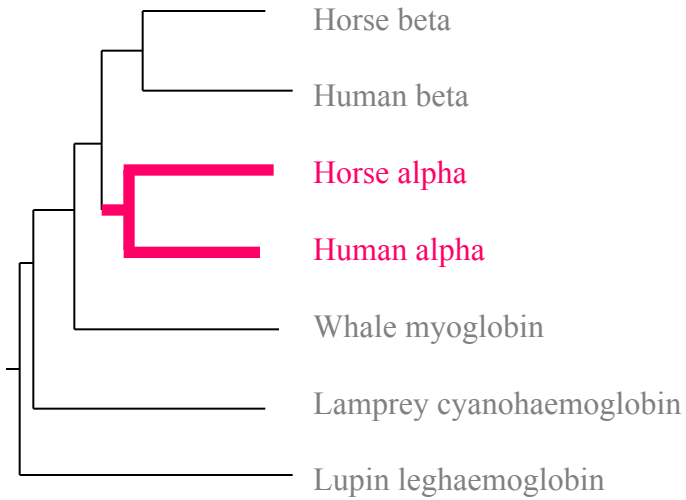
"Guide Tree"



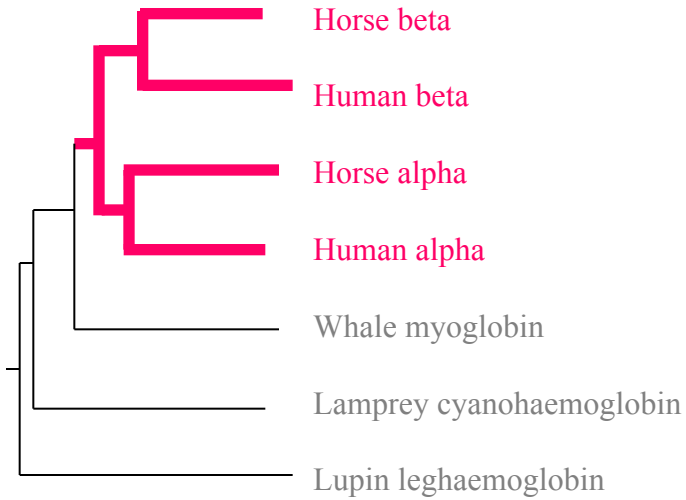
```
Human beta          --------VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLST
Horse beta          --------VQLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDLSN
Human alpha         ---------VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-
Horse alpha         ---------VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF-DLS-
Whale myoglobin     ---------VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKT
Lamprey globin      PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLTT
Lupin globin        --------GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE
                              *:  :   :   *  .           :  .:   * :   *  :   .

Human beta          PDAVMGNPKVKAHGKKVLGAFSDGLAHLDN-----LKGTFATLSELHCDKLHVDPENFRL
Horse beta          PGAVMGNPKVKAHGKKVLHSFGEGVHHLDN-----LKGTFAALSELHCDKLHVDPENFRL
Human alpha         ----HGSAQVKGHGKKVADALTNAVAHVDD-----MPNALSALSDLHAHKLRVDPVNFKL
Horse alpha         ----HGSAQVKAHGKKVGDALTLAVGHLDD-----LPGALSNLSDLHAHKLRVDPVNFKL
Whale myoglobin     EAEMKASEDLKKHGVTVLTALGAILKKKGH-----HEAELKPLAQSHATKHKIPIKYLEF
Lamprey globin      ADQLKKSADVRWHAERIINAVNDAVASMDDT--EKMSMKLRDLSGKHAKSFQVDPQYFKV
Lupin globin        VP--QNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKGVAD-AHFPV
                          . .:: *.  :   .                  :  *.  *  .       : .

Human beta          LGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------
Horse beta          LGNVLVVVLARHFGKDFTPELQASYQKVVAGVANALAHKYH------
Human alpha         LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
Horse alpha         LSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTSKYR------
Whale myoglobin     ISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
Lamprey globin      LAAVIADTVAAG---D------AGFEKLMSMICILLRSAY-------
Lupin globin        VKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
                    :   :  .:      .      ..       .   :
```



```
Human beta        --------VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLST
Horse beta        --------VQLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDLSN
Human alpha       ---------VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-
Horse alpha       ---------VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF-DLS-
Whale myoglobin   ---------VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKT
Lamprey globin    PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLTT
Lupin globin      --------GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE
                            *:  :   :   *  .           :  .:   * :   *  :   .

Human beta        PDAVMGNPKVKAHGKKVLGAFSDGLAHLDN-----LKGTFATLSELHCDKLHVDPENFRL
Horse beta        PGAVMGNPKVKAHGKKVLHSFGEGVHHLDN-----LKGTFAALSELHCDKLHVDPENFRL
Human alpha       ----HGSAQVKGHGKKVADALTNAVAHVDD-----MPNALSALSDLHAHKLRVDPVNFKL
Horse alpha       ----HGSAQVKAHGKKVGDALTLAVGHLDD-----LPGALSNLSDLHAHKLRVDPVNFKL
Whale myoglobin   EAEMKASEDLKKHGVTVLTALGAILKKKGH-----HEAELKPLAQSHATKHKIPIKYLEF
Lamprey globin    ADQLKKSADVRWHAERIINAVNDAVASMDDT--EKMSMKLRDLSGKHAKSFQVDPQYFKV
Lupin globin      VP--QNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKGVAD-AHFPV
                        . .:: *.  :   .                  :  *.  *  .       : .

Human beta        LGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------
Horse beta        LGNVLVVVLARHFGKDFTPELQASYQKVVAGVANALAHKYH------
Human alpha       LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
Horse alpha       LSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTSKYR------
Whale myoglobin   ISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
Lamprey globin    LAAVIADTVAAG---D------AGFEKLMSMICILLRSAY-------
Lupin globin      VKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
                  :   :  .:      .      ..       .   :
```



```
Human beta          --------VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLST
Horse beta          --------VQLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDLSN
Human alpha         ---------VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-
Horse alpha         ---------VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF-DLS-
Whale myoglobin     ---------VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKT
Lamprey globin      PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLTT
Lupin globin        --------GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE
                              *:  :   :   *  .           :  .:   * :   *  :   .

Human beta          PDAVMGNPKVKAHGKKVLGAFSDGLAHLDN-----LKGTFATLSELHCDKLHVDPENFRL
Horse beta          PGAVMGNPKVKAHGKKVLHSFGEGVHHLDN-----LKGTFAALSELHCDKLHVDPENFRL
Human alpha         ----HGSAQVKGHGKKVADALTNAVAHVDD-----MPNALSALSDLHAHKLRVDPVNFKL
Horse alpha         ----HGSAQVKAHGKKVGDALTLAVGHLDD-----LPGALSNLSDLHAHKLRVDPVNFKL
Whale myoglobin     EAEMKASEDLKKHGVTVLTALGAILKKKGH-----HEAELKPLAQSHATKHKIPIKYLEF
Lamprey globin      ADQLKKSADVRWHAERIINAVNDAVASMDDT--EKMSMKLRDLSGKHAKSFQVDPQYFKV
Lupin globin        VP--QNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKGVAD-AHFPV
                          . .:: *.  :   .                  :  *.  *  .       : .

Human beta          LGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------
Horse beta          LGNVLVVVLARHFGKDFTPELQASYQKVVAGVANALAHKYH------
Human alpha         LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
Horse alpha         LSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTSKYR------
Whale myoglobin     ISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
Lamprey globin      LAAVIADTVAAG---D------AGFEKLMSMICILLRSAY-------
Lupin globin        VKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
                    :   :  .:      .      ..       .   :
```

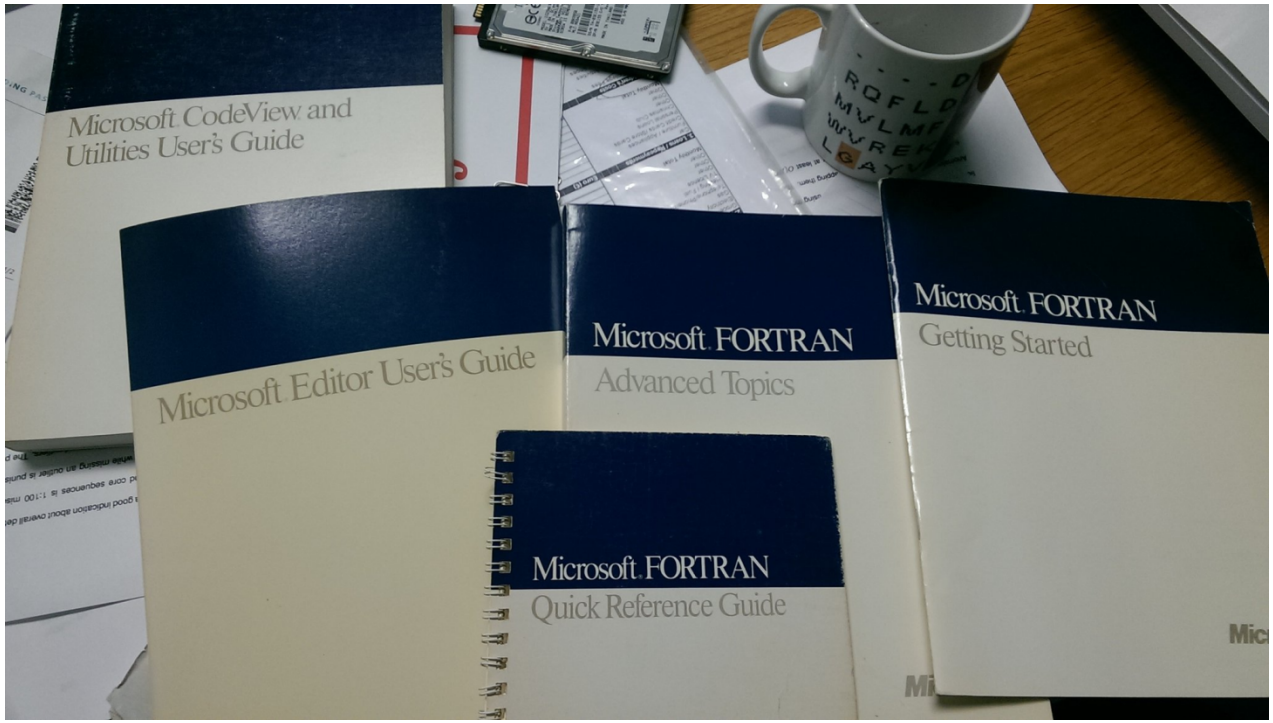


1988!

- Paul Sharp lab.
  - TCD
- No internet (much)
- IBM PC
  - MS-DOS
  - 640kb memory
  - 0-20MB hard disk

Fortran!

- MS Fortran for MS-DOS
- No recursion
- No memory management (much)



Clustal 1-3

- Fast k-tuple pairwise alignments (Clustal 1)
- UPGMA (Clustal 2)
- Align larger and larger alignments (Clustal 3)
  - Fast approximate word based alignments

ELSEVIER

Gene

Volume 73, Issue 1, 15 December 1988, Pages 237-244

CLUSTAL: a package for performing multiple sequence alignment on a microcomputer

Desmond G. Higgins, Paul M. Sharp

Show more

https://doi.org/10.1016/0378-1119(88)90330-7

Abstract

An approach for performing multiple alignments of large numbers of amino acid or nucleotide sequences is described. The method is based on first deriving a phylogenetic tree from a matrix of all pairwise sequence similarity scores, obtained using a fast pairwise alignment algorithm. Then the multiple alignment is achieved from a series of pairwise alignments of clusters of sequences, following the order of branching in the tree. The method is sufficiently fast and economical with memory to be easily implemented on a microcomputer, and yet the results obtained are comparable to those from packages requiring mainframe computer facilities.

Previous article in issue

Next article in issue

Keywords

Cluster analysis; phylogenetic tree; protein secondary structure; RNA secondary structure; globin; 5S RNA; dendrogram

Abbreviations

aa, amino acid(s); ASCII, American Standard Code for Information Interc
nucleotide(s); S, sedimentation constant; UPGMA, Unweighted Pair-Grou
Arithmetic Averages

2-Seq Alignment?

- $O(L^2)$
- L = 1000?
  - 1-4 MB memory
- Myers and Miller,1988

CABIOS
Vol.4. no.1. 1988
Pages 11 – 17

Optimal alignments in linear space

Eugene W.Myers[1,2] and Webb Miller[2]

Abstract

Space, not time, is often the limiting factor when computing optimal sequence alignments, and a number of recent papers in the biology literature have proposed space-saving strategies. However, a 1975 computer science paper by Hirschberg presented a method that is superior to the new proposals, both in theory and in practice. The goal of this paper is to give Hirschberg's idea the visibility it deserves by developing a linear-space version of Gotoh's algorithm, which accommodates affine gap penalties. A portable C-software package implementing this algorithm is available on the BIONET free of charge.

Introduction

Consider the following problem. Given sequences $A = a_1a_2...a_M$ and $B = b_1b_2...b_N$, find a set of 'evolutionary operations' that converts A to B and minimizes the sum of the operations' costs. The allowed operations are (i) replace one symbol by another, (ii) delete k consecutive symbols, or (iii) insert k consecutive symbols. In addition, the problem statement requires that every symbol of A must be either replaced or deleted. Replacement costs are specified by a table, w, where $w(a,b)$ gives the cost of replacing a by b. Note that a symbol of A is effectively left unedited if it is replaced by itself at no cost, i.e. $w(a,a) = 0$. Two non-negative constants, g and h, specify an affine function, $gap(k) = g + hk$, for the cost of a k-symbol indel (insertion or deletion). Informally, opening up a gap costs g and each symbol in the gap costs h.

The problem is often formulated as maximizing the similarity score of an alignment, rather than minimizing the difference score of a conversion. A bonus $\sigma(a,b)$ is added for every aligned pair (a,b) and a 'gap penalty' $q + rk$ is subtracted for every k-symbol gap. This formulation is converted to a difference problem by the transformations

$$w(a,b) = \sigma_{max} - \sigma(a,b) \text{ for all pairs } (a,b)$$
$$g = q$$
$$h = r + \tfrac{1}{2}\sigma_{max}$$

where $\sigma_{max} = \max_{(a,b)}\sigma(a,b)$ (Smith *et al.*, 1981). Thus, to produce an alignment that maximizes the similarity score, first apply these transformations and then run the program described in this paper with the resulting w, g and h. If the minimum conversion score is C, then the corresponding maximum alignment score is $\tfrac{1}{2}(M + N)\sigma_{max} - C$.

Gotoh (1982) gave an algorithm that solves such problems in $O(MN)$ time. If only the minimum cost is desired, then it is easy to implement the algorithm in $O(N)$ space, where N can be taken as the shorter sequence length. If one also desires a set of operations attaining the minimum cost, then straightforward implementations need $O(MN)$ space. In practice, this space requirement often limits the method's applicability, and several papers (Taylor, 1984; Watanabe *et al.*, 1985; Altschul and Erickson, 1986; Gotoh, 1986, 1987) have presented strategies that reduce space consumption by constant factors. These papers fail to note that Hirschberg (1975) showed how to produce an optimal conversion or alignment in $O(N)$ space. When only a single optimal alignment of A and B is desired, Hirschberg's approach is superior to the others. For example, in one megabyte of memory, our program based on Hirschberg's method can align two sequences of length 62 500. Altschul and Erickson (1986) propose keeping 7 bits for each of MN entries, so the limit for their method is $7N^2 \leq 8 \times 10^6$, or $N < 1070$. Moreover, any program that packs and unpacks bits or uses disk storage is doomed to be slow and, probably, non-portable.

$O(MN)$-space methods permit the construction of all optimal alignments. However, the number of alignments that attain the minimum cost is often astronomical, in part because a brute force enumeration lists many arrangements whose differences are insignificant to the user. Moreover, when one is searching for a particular 'biologically meaningful' arrangement, it may be necessary to consider slightly sub-optimal alignments (Waterman, 1983; Waterman and Byers, 1985). One alternative to explicitly constructing all optimal alignments is to modify our linear-space program to produce 'left-most' and 'right-most' optimal alignments that delineate the range of possibilities. In any case, it is important to understand that a single optimal alignment can be found in far less space than is needed to record 'traceback' information for finding all optimal alignments.

Hirschberg's original presentation treats a simpler alignment problem, known as the longest common subsequence problem, where $w(a,b) = 1$ if $a \neq b$, $w(a,a) = 0$, and $gap(k) = k$. However, the approach is quite general. To the best of our

[1]*Department of Computer Science, University of Arizona, Tucson, AZ 85721, USA*

[2]*Department of Computer Science, The Pennsylvania State University, University Park, PA 16802, USA*

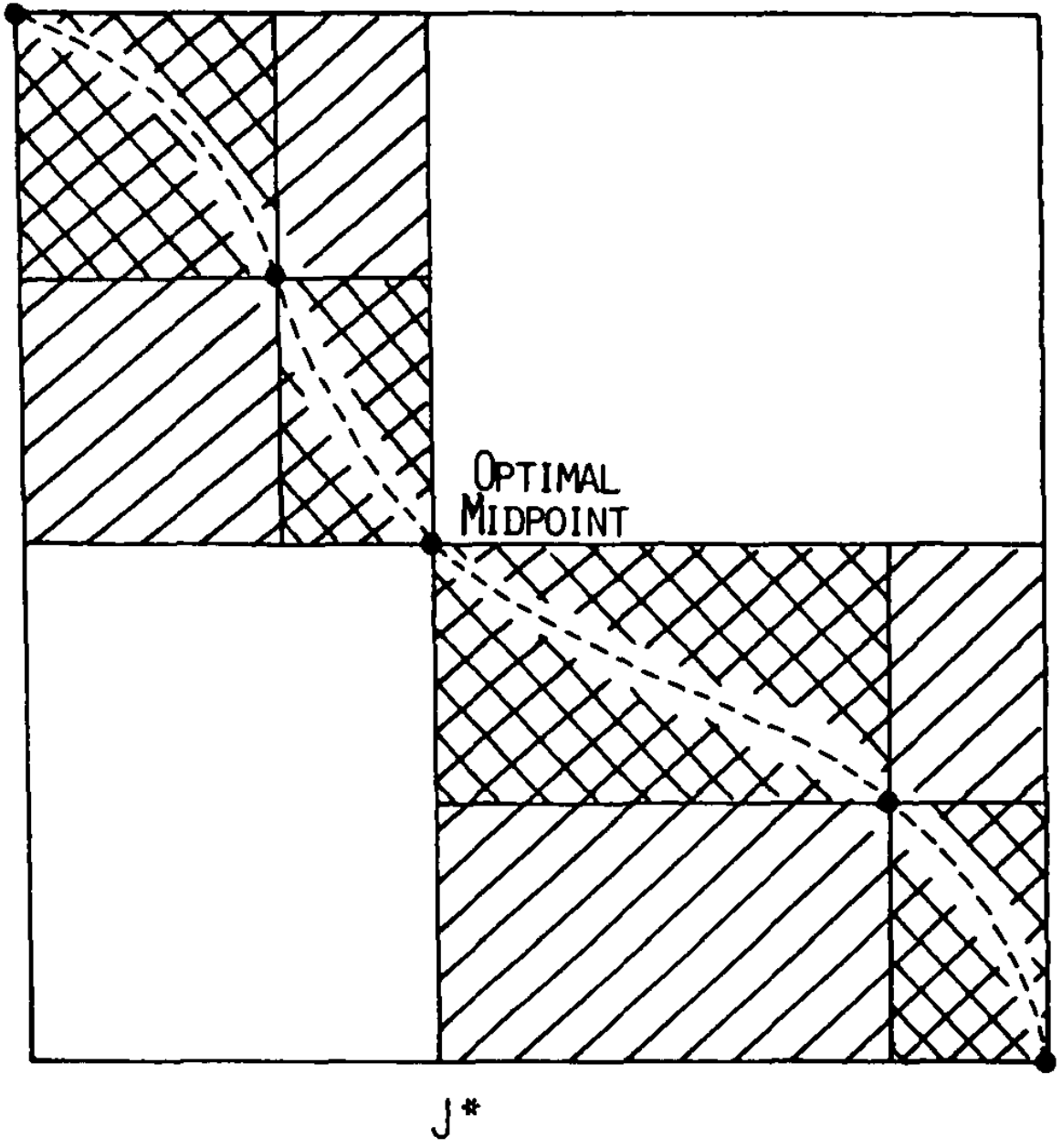


Fig. 2. Splitting the problem into sub-problems.

Clustal 4

- Full progressive alignment
- On PC
- Fast
- Accurate

CABIOS COMMUNICATIONS

Vol.5. no.2. 1989
Page 151 - 153

Fast and sensitive multiple sequence alignments on a microcomputer

Desmond G.Higgins and Paul M.Sharp*

Abstract

A strategy is described for the rapid alignment of many long nucleic acid or protein sequences on a microcomputer. The program described can handle up to 100 sequences of 1200 residues each. The approach is based on progressively aligning sequences according to the branching order in an initial phylogenetic tree. The results obtained using the package appear to be as sensitive as those from any other available method.

Introduction

In the recent literature on biological sequence analysis, at least a dozen methods for performing multiple alignments of nucleic acid or protein sequences have been described [e.g. Bains (1986), Sobel and Martinez (1986), Barton and Sternberg (1987), Feng and Doolittle (1987), Santibanez and Rohde (1987), Taylor (1987)]. The motivation for this effort has been the need for the automatic alignment of three or more sequences for the purposes of evolutionary or structural comparisons or for attempting to demonstrate similarity between sets of sequences. In this paper, we describe a strategy which we believe offers the best combination of speed and sensitivity available for any multiple alignment method. We offer a program which can perform multiple alignments of up to 100 sequences of maximum length 1200 residues on a microcomputer in a reasonable amount of time. We judge the program to be

for each residue in one set compared against each residue in the second set. Any gaps introduced into either set of sequences are scored as single gaps. The main difficulty in using this approach on a microcomputer arises from the excessive memory requirements of the Needleman and Wunsch (1970) method—memory usage is proportional to the square of the average sequence length.

In a previous paper (Higgins and Sharp, 1988) we described a strategy for the very rapid multiple alignment of large numbers of sequences on a microcomputer. This method also comprised a progressive approach, using the fast, but approximate, two-sequence alignment method of Wilbur and Lipman (1983). While this approach is extremely rapid and economical with core memory, it works well only for closely related sequences. We did not consider using the exactly optimal method of Needleman and Wunsch (1970) for the progressive alignments because of the excessive memory requirements. However, a recent paper by Myers and Miller (1988) demonstrates how to achieve exactly optimal alignments of two sequences where memory usage varies only linearly with sequence length, without making use of bit packing or secondary disk storage. Thus, a progressive series of alignments of larger and larger groups of sequences, using the method of Myers and Miller (1988) for each alignment, is the key to the current approach.

System

Distribution?

- Mainly by post
- The EMBL Mail Server

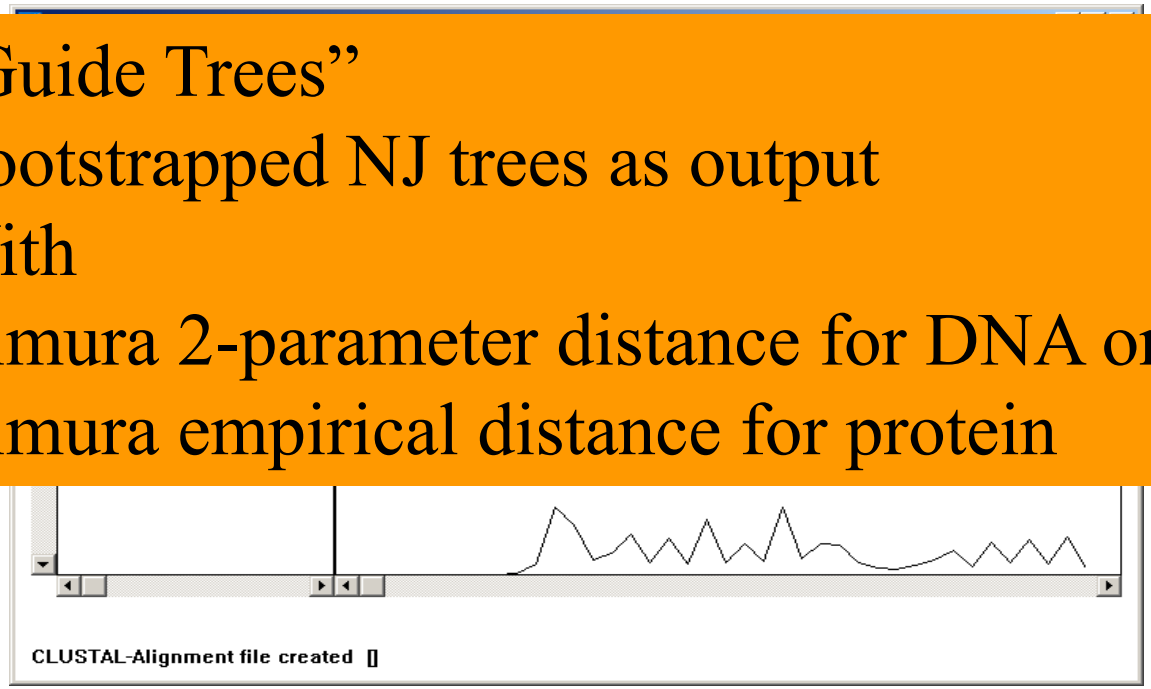
System

The program described in this paper was written in standard FORTRAN 77 and compiled using the Microsoft FORTRAN compiler, version 4.0. Program performance was tested on an IBM AT compatible microcomputer, running at 10 MHz with no maths coprocessor, 640 kbytes of memory and a hard disk. This program (CLUSTAL4) is an extension to the package described in Higgins and Sharp (1988). Copies of the executable files, documentation and test data files will be sent on request. Please send three 5.25 inch floppies formatted to 360 kbytes, or one high density 5.25 inch floppy formatted to 1.2 Mbytes.

Clustal

- Clustal1-Clustal4
  - 1988, Paul Sharp, Dub
- Clustal V 1992
  - EMBL Heidelberg,
  - Rainer Fuchs
  - Alan Bleasby
- Clustal W, Clustal X 1994-2007
  - EBI, University College Cork
  - Toby Gibson, EMBL, Heidelberg
  - Julie Thompson, ICGEB, Strasbourg
- Clustal W and Clustal X 2.0 2007
  - University College Dublin
- **>130k citations on Google Scholar**

"Guide Trees"
Bootstrapped NJ trees as output
With
Kimura 2-parameter distance for DNA or
Kimura empirical distance for protein



science foundation ireland
fondúireacht eolaíochta éireann

www.clustal.org

LOWRY, OH; ROSEBROUGH, NJ; FARR, AL; RANDAL
LAEMMLI, UK
BRADFORD, MM
SANGER, F; NICKLEN, S; COULSON, AR
CHOMCZYNSKI, P; SACCHI, N
TOWBIN, H; STAEHELIN, T; GORDON, J
LEE, CT; YANG, WT; PARR, RG
FOLCH, J; LEES, M; STANLEY, GHS
BECKE, AD
THOMPSON, JD; HIGGINS, DG; GIBSON, TJ
KAPLAN, EL; MEIER, P
ALTSCHUL, SF; GISH, W; MILLER, W; MYERS, EW; L
Sheldrick, George M.
MURASHIGE, T; SKOOG, F
Altschul, SF; Madden, TL; Schaffer, AA; Zhang, JH; Zha
FOLSTEIN, MF; FOLSTEIN, SE; MCHUGH, PR
Perdew, JP; Burke, K; Ernzerhof, M
SOUTHERN, EM
BLIGH, EG; DYER, WJ
SAITOU, N; NEI, M
COX, DR
SHANNON, RD
Otwinowski, Z; Minor, W
Livak, KJ; Schmittgen, TD
BECKE, AD
DUBOIS, M; GILLES, KA; HAMILTON, JK; REBERS, PA
REYNOLDS, ES
WEBER, K; OSBORN, M
Thompson, JD; Gibson, TJ; Plewniak, F; Jeanmougin, F
CHIRGWIN, JM; PRZYBYLA, AE; MACDONALD, RJ; R

http://www.nature.com/news/the-top-100-papers-1.16224

Top 30 papers on ISI up to 2015

Since Clustal…

- >100 new programs in past 15 years
- 1999: BENCHMAR
- T-Coffee (and R-Coffee, M-Coffee, 3D-Coffee)
  - Cedric Notredame
    http://www.tcoffee.org
- MAFFT
  - Kazutaka Katoh, 2001
    - http://mafft.cbrc.jp/alignment/software/
- MUSCLE
  - Bob Edgar, ISMB 2004
    - http://www.drive5.com/muscle
- PROBCONS
  - Tom Do, Michael Brudno, Serafim Batzoglou, ISMB 2004

Balibase, Oxbench, Prefab, Sabre

Consistency!

MSAProbs
Yongchao Liu, Bertil Schmidt, Douglas L. Maskell (2010)
"MSAProbs: multiple sequence alignment based on pair
hidden Markov models and partition function posterior probabilities".
Bioinformatics, 26(16): 1958 -1964

Two kinds of programs

- Fast (<10,000 seqs)
  - Clustal W
  - MAFFT
    - mafft --partree >>10,000 seqs
  - MUSCLE
  - Kalign
- Accurate (100s seqs)
  - T-Coffee
  - ProbCons
  - MAFFT
  - MSAProbs

Prank and Pagan

Accurate?

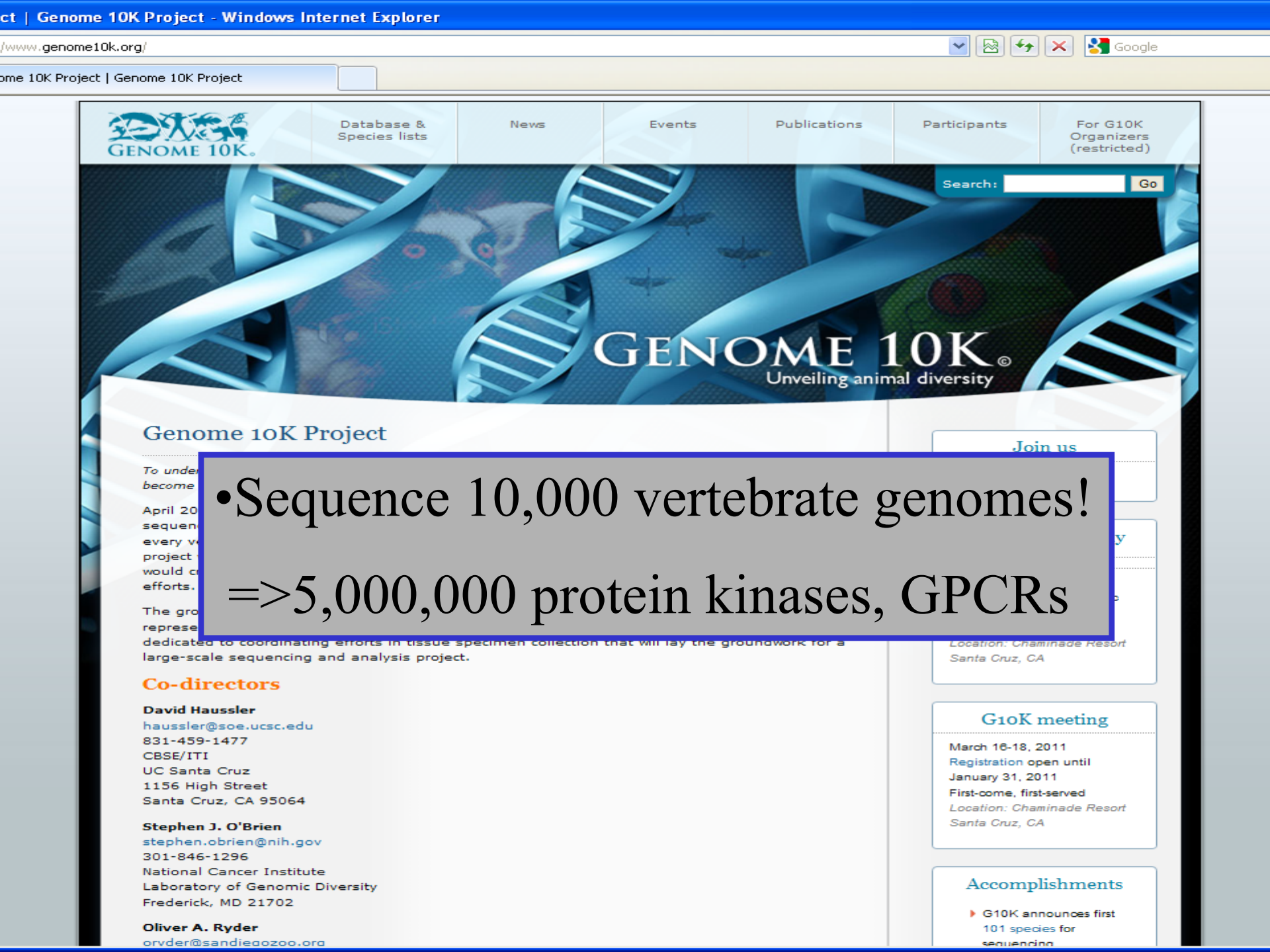
- Protein structure comparisons
- Simulation

- Protein structure versus phylogeny

- Accurate analysis of indels/closely related sequences/analysing sites and selection?
  - Use e.g. Prank

So: What is the Problem?

- What if N >> 100,000?
- e.g. SSU rRNA
  - www.arb-silva.de
  - **4,346,367** seqs
- e.g. ABC transporters
  - PFAM May 2015
  - ABC_tran PF00005
  - **150k** seqs
- Metagenomics

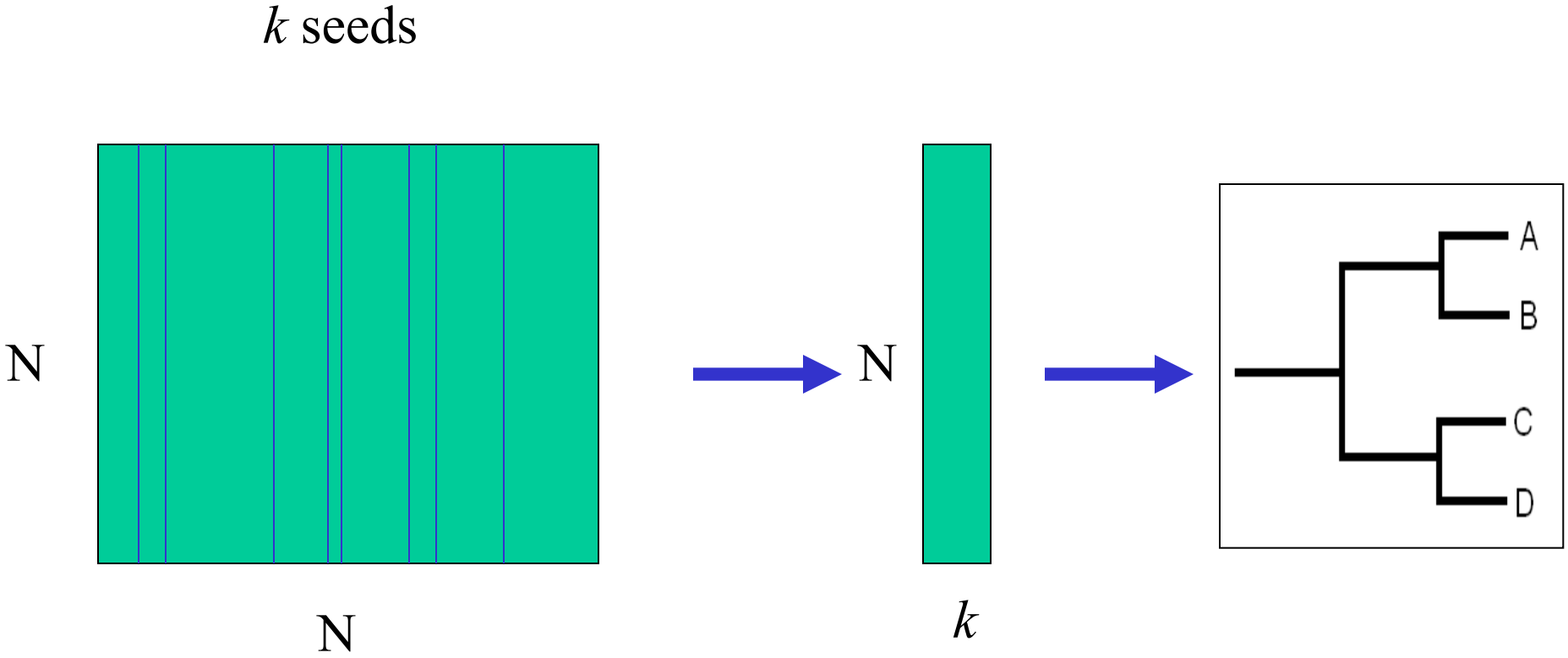

ct | Genome 10K Project - Windows Internet Explorer
/www.genome10k.org/
Google
ome 10K Project | Genome 10K Project
GENOME 10K
Database & Species lists
News
Events
Publications
Participants
For G10K Organizers (restricted)
Search:
Go
GENOME 10K
Unveiling animal diversity
Genome 10K Project
•Sequence 10,000 vertebrate genomes!
=>5,000,000 protein kinases, GPCRs
large-scale sequencing and analysis project.
Co-directors
David Haussler
haussler@soe.ucsc.edu
831-459-1477
CBSE/ITI
UC Santa Cruz
1156 High Street
Santa Cruz, CA 95064
Stephen J. O'Brien
stephen.obrien@nih.gov
301-846-1296
National Cancer Institute
Laboratory of Genomic Diversity
Frederick, MD 21702
Oliver A. Ryder
Join us
Location: Chaminade Resort
Santa Cruz, CA
G10K meeting
March 16-18, 2011
Registration open until
January 31, 2011
First-come, first-served
Location: Chaminade Resort
Santa Cruz, CA
Accomplishments
G10K announces first 101 species for


sfi
CLUSTAL
CLUSTAL

Big Alignments?

- Once a guide tree is made:
  - Multiple alignment: for N seqs, fixed len L
  - Time $O(N)$
- To make a guide tree:
  - Time and memory: $O(N^2)$
- Maximum alignment size 10-20k seqs.

mBED



Blackshields G, Sievers F, Shi W, Wilm A, Higgins DG. (2010)
Sequence embedding for fast construction of guide trees for multiple sequence alignment.
Algorithms Mol Biol. 14;5:21.

Clustal Ω

- Released April 2011
- Scalable
  - mBed
  - Blackshields, et al. (2010) Algorithms in Molecular Biology.
- Accurate
  - Hidden Markov models (HMMs)
  - HHalign
  - Johannes Söding, Munich.
- Re-use old alignments
  - Kevin Karplus
  - UCSC

Molecular Systems Biology 7; Article number 539; doi:10.1038/msb.2011.75
Citation: *Molecular Systems Biology* 7: 539

www.molecularsystemsbiology.com

molecular systems biology

REPORT

Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega

Fabian Sievers[1,8], Andreas Wilm[2,8], David Dineen[1], Toby J Gibson[3], Kevin Karplus[4], Weizhong Li[5], Rodrigo Lopez[5], Hamish McWilliam[5], Michael Remmert[6], Johannes Söding[6], Julie D Thompson[7] and Desmond G Higgins[1,*]

[1] School of Medicine and Medical Science, UCD Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Dublin, Ireland, [2] Computational and Systems Biology, Genome Institute of Singapore, Singapore, [3] Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany, [4] Department of Biomolecular Engineering, University of California, Santa Cruz, CA, USA, [5] EMBL Outstation—European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK, [6] Gene Center Munich, University of Munich (LMU), Muenchen, Germany and [7] Département de Biologie Structurale et Génomique, IGBMC (Institut de Génétique et de Biologie Moléculaire et Cellulaire), CNRS/INSERM/Université de Strasbourg, Illkirch, France

[8] These authors contributed equally to this work

* Corresponding author. UCD Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Belfield, Dublin 4, Ireland. Tel.: + 353 1 716 6833; Fax: + 353 1 716 6713; E-mail: des.higgins@ucd.ie

Received 23.7.11; accepted 6.9.11

- Alig
  - 6
- Mor
  - M
  - A
- Dow
  - w

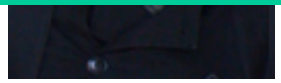
BaliBase	% correct	time(s)
Clustal Omega	55.4	539
Clustal W	37.4	766
Mafft (default)	45.8	68
Muscle	47.5	789
Kalign	50.1	21
T-Coffee	55.1	81041
Probcons	55.8	13086
Mafft (auto/consistency)	58.8	1475
MsaProbs	60.7	12382

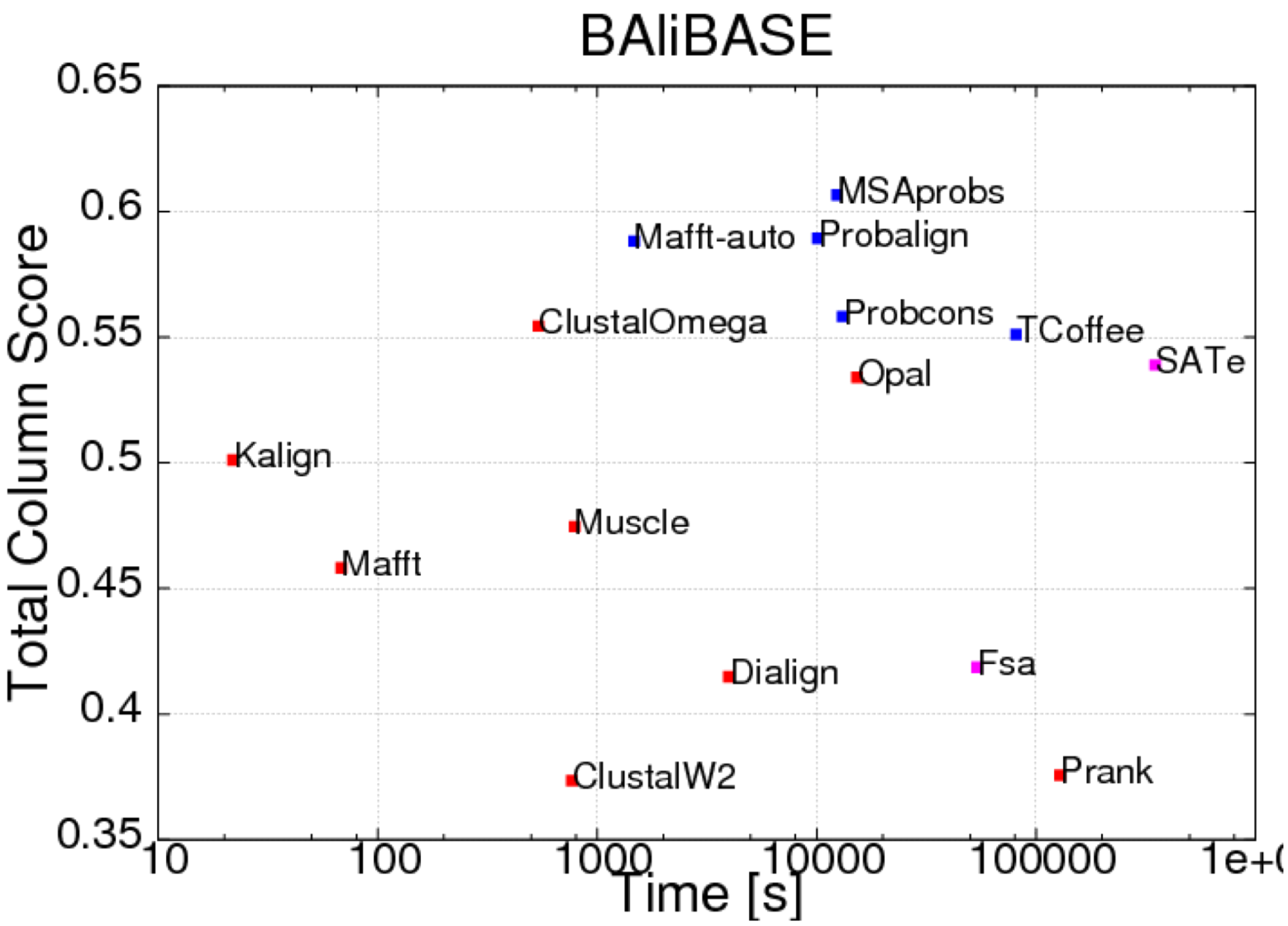
arzl

vers

ilm

een




BAliBASE
Total Column Score
0.65
0.6
0.55
0.5
0.45
0.4
0.35
10
100
1000
10000
100000
1e+(
Time [s]
MSAprobs
Mafft-auto
Probalign
ClustalOmega
Probcons
TCoffee
SATe
Opal
Kalign
Muscle
Mafft
Dialign
Fsa
ClustalW2
Prank

Clustal Omega Servers

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

- Download your own copy
  - Command-line only (Unix/Linux style)
    - Visualise with Jalview or SeaView (Manolo Gouy)
  - Proteins or DNA/RNA

– www.clustal.org

Clustal

Paul Sharp, Edinburgh
Rainer Fuchs, EMBL
Alan Bleasby, Daresbury
Toby Gibson, Ramu Chenna, Nigel Brown, EMBL
Julie Thompson, Francois Jeanmougin, Fred Plewniak, Strasbourg
Paul Mcgettigan, Mark Larkin, Andreas Wilm, Fabian Sievers, UCD

mBED

Gordon Blackshields
Mark Larkin

Clustal Omega

Fabian Sievers, Andreas Wilm, David Dineen, UCD
Johannes Soeding, Michael Remmert, Munich
Rodrigo Lopez, Hamish MacWilliam, Weizhong Li, EBI
Kevin Karplus, UCSC

Current Alignment Group

Fabian Sievers, Quan Le
Kieran Boyce, Gearóid Fox and Peter Jehl

sfi
science foundation ireland
fondúireacht eolaíochta éireann