Introduction to Computational Biology; An Evolutionary Approach: Optimal Pairwise Alignment

Bernhard Haubold & Thomas Wiehe

ICB © 2006 Birkhäuser Verlag

Outline

1 Pairwise Alignment

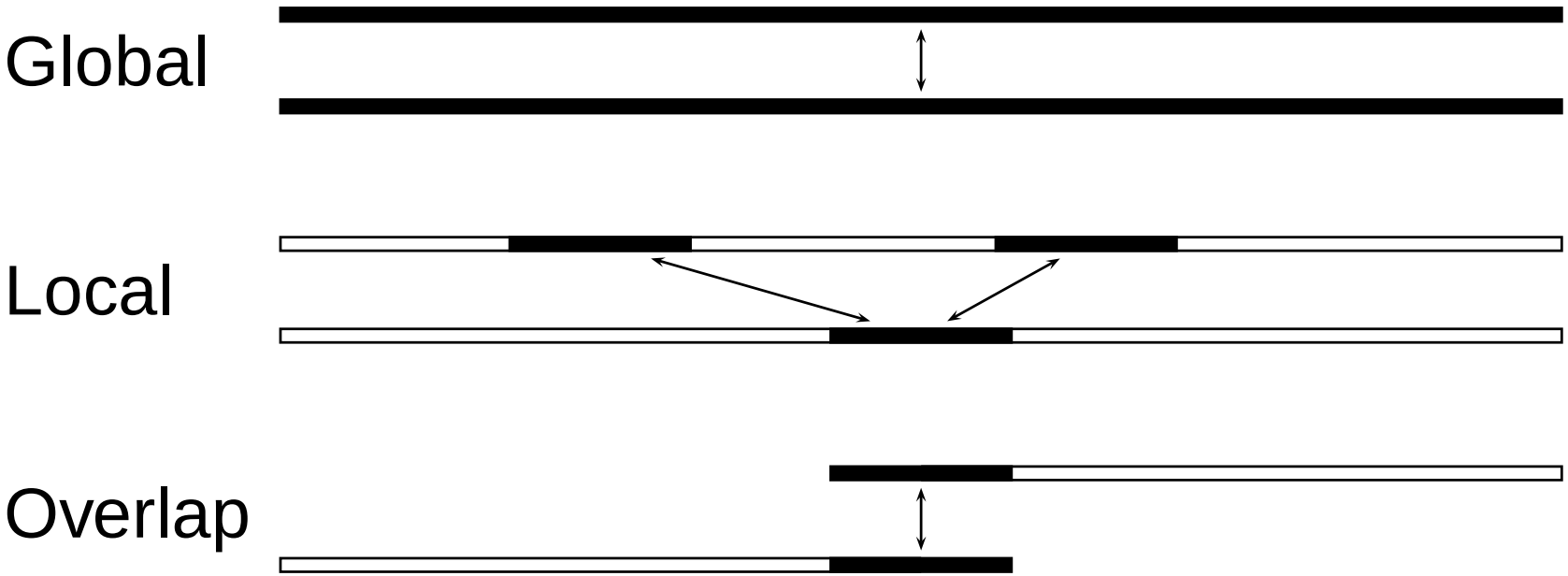
2 Amino Acid Substitution Matrices

3 Number of Alignments

4 Global Alignment

ICB © 2006 Birkhäuser Verlag

Types of Pairwise Alignments



Hox Alignment

```
D  MMDVSSMYGNHPHHHHPHANAYDGYSTTTASAANASSYFAPQQHQPHLQLQQQQQ   55
H  -MDNARMN---------SFLEYPILSSGDSGTCSARAYPS---------------   30

D  HQHLQQPQQHLTYNGYESSSPGNYYPQQQAQLTPPPTSSHQVVQQHQQQQQAQQQ  110
H  -DHRITTFQSCAVSANSCGGDDRFLVGRGVQIGSP--------------------   64

D  QLYPHSHLFSPSAAEYGITTSTTTGNPGTPLHPSSHSPADSYYESDSVHSYYATA  165
H  HHHHHHHHHHPQPATY---Q--TSGNLGVSYSHSSCGPS---YGSQNFSAPYSPY  111

D  AVATVAPPSNSSPITAANASATSNTQQQQQQAAIISSENGMMYTNLDCMYPTAQA  220
H  ALN-----------QEADVSGG----------------------------YPQC--  126

D  QAPVHGYAGQIEEKYAAVLHASYAPGMVLEDQDPMMQQATQSQMWHHQQHLAGSY  275
H  -AP-------------------------------------------------AVY  131

D  ALDAMDSLGMHAHMHHGLPHGHLGNLANNPHQQQPQVQQQQQQPHQQPQHPQNQS  330
H  SGNLSSPMVQHHHHHQGYAGGAVG--------------------SPQYIHHSYG--  165

D  PAAHQQHHQNSVSPNGGMNRQQRGGVISPGSSTSSSTSASNGAHPASTQSKSPNH  385
H  ----QEHQSLALATYN-----------------N-SLSPLHASHQEACRSPASET  198

D  SSSIPTYKWMQLKRNVPKPQAPSYLPAPKLPASGIASMHDYQMNGQLDMCRGGGG  440
H  SSPAQTFDWMKVKRNPPK-------------------------------------  216
```

Development of Alignment Algorithms

Year	Authors	Algorithm	Approach
1970	Needleman & Wunsch	Global	Optimal
1981	Smith & Waterman	Local	Optimal
1988	Pearson & Lipman	FASTA	Heuristic
1990	Altschul *et al.*	BLAST	Heuristic
1997	Altschul *et al.*	Gapped BLAST	Heuristic

ICB © 2006 Birkhäuser Verlag

Possible Alignments

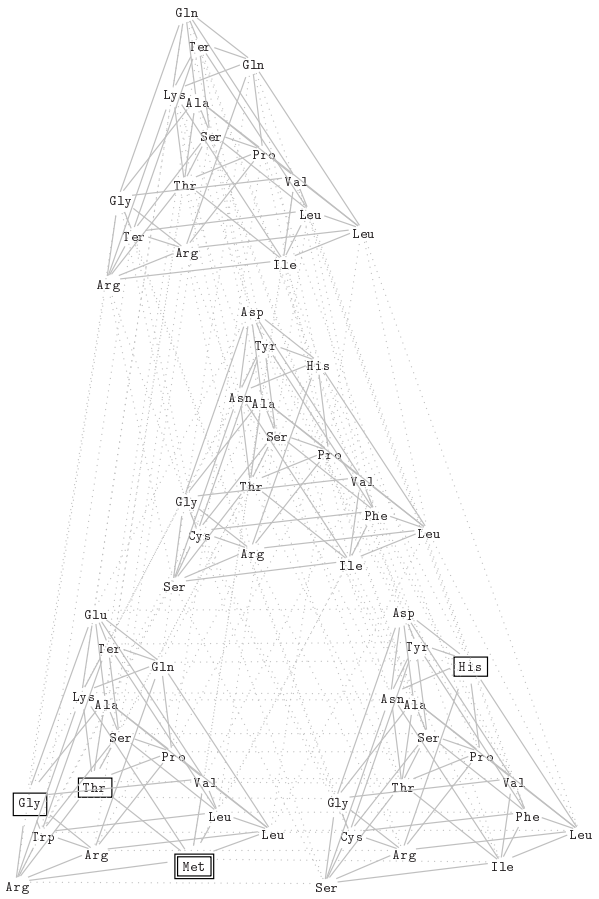
$S_1 = \texttt{AACGT}$, $S_2 = \texttt{ACCGTT}$

1	2	3	4	5	6
AACGT- ACCGTT	-AACGT ACCGTT	A-ACGT ACCGTT	AACGT---- ---ACCGTT	AA-CGT- A-CCGTT	-A-A-C-G-T- A-C-C-G-T-T

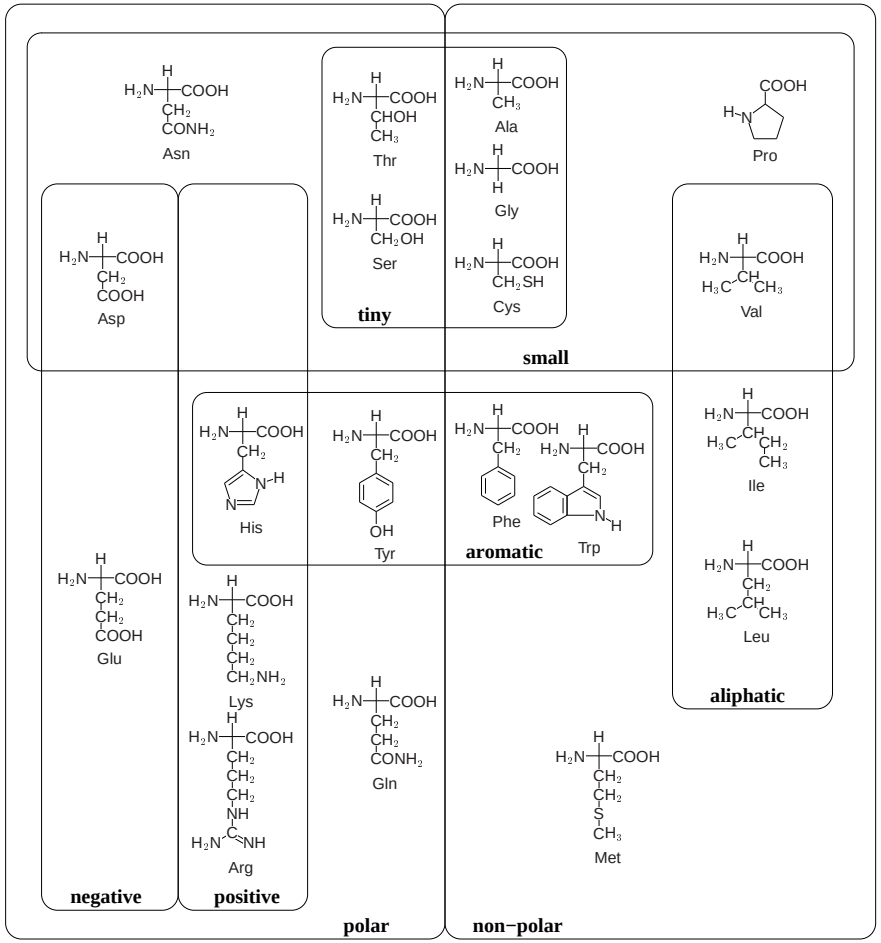
Scoring

$$
\begin{aligned}
G &= l \times g_{\mathrm{e}} \\
G &= g_{\mathrm{o}} + l \times g_{\mathrm{e}}
\end{aligned}
$$

Amino Acids in Sequence Space



Structures of Amion Acids

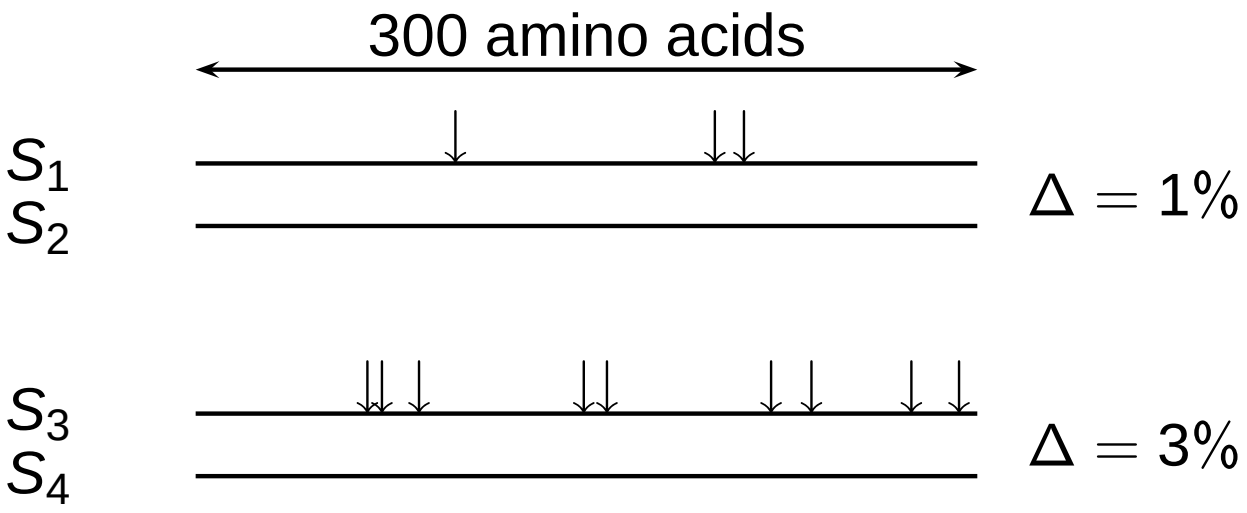


Amino Acid Mutation Probabilities

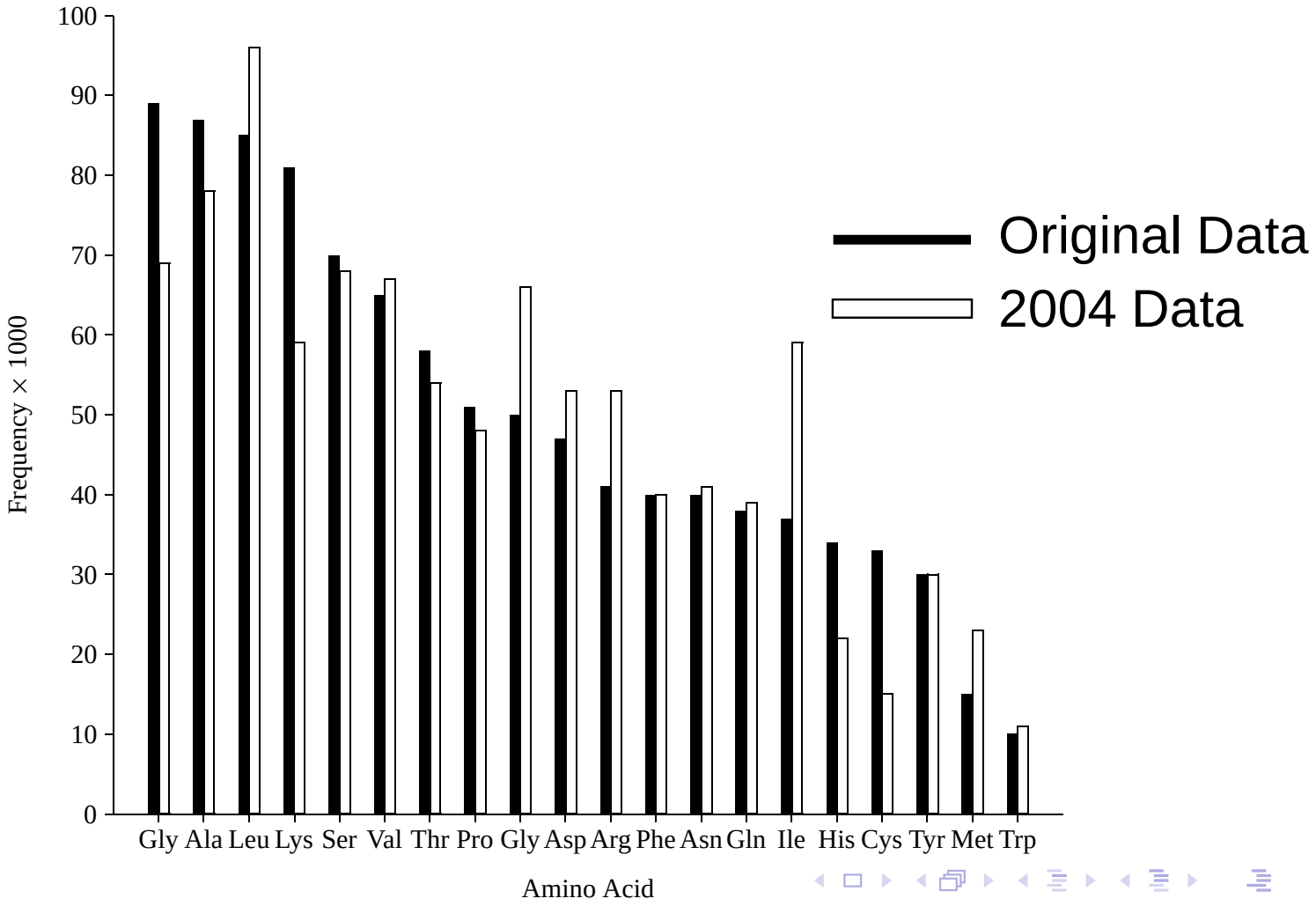
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C	1	1	0	0	9973	0	0	0	1	1	0	0	1	0	0	5	1	0	3	2
Q	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
E	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
G	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
L	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
K	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
F	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
P	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
S	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
T	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Y	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
V	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

ICB © 2006 Birkhäuser Verlag

Normalization of Mutation Probabilities



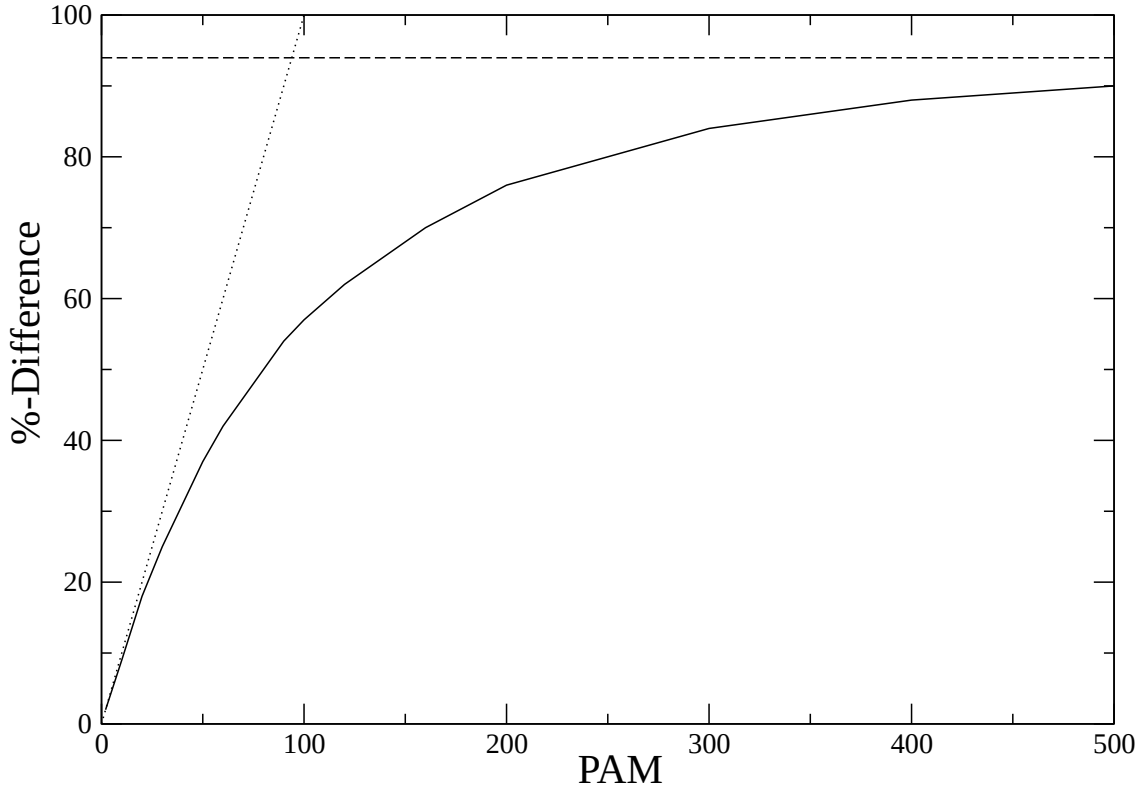
Amino Acid Background Frequencies



%-Difference

$$D = \left(1 - \sum_{i}^{20} M_{ii} \times f_i\right) \times 100,$$

PAM Number vs. %-Difference



Derivation of PAM Matrix

$$
\begin{aligned}
R_{ij} &= \frac{M_{ij}}{f_i} \\
s_{ij} &= \log_2 \left(R_{ij} \right)
\end{aligned}
$$

ICB © 2006 Birkhäuser Verlag

PAM160

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	**2**	-2	0	0	-2	-1	0	1	-2	-1	-2	-2	-1	-3	1	1	1	-5	-3	0
R	-2	**6**	-1	-2	-3	1	-2	-3	1	-2	-3	3	-1	-4	-1	-1	-1	1	-4	-3
N	0	-1	**3**	2	-4	0	1	0	2	-2	-3	1	-2	-3	-1	1	0	-4	-2	-2
D	0	-2	2	**4**	-5	1	3	0	0	-3	-4	0	-3	-6	-2	0	-1	-6	-4	-3
C	-2	-3	-4	-5	**9**	-5	-5	-3	-3	-2	-6	-5	-5	-5	-3	0	-2	-7	0	-2
Q	-1	1	0	1	-5	**5**	2	-2	2	-2	-2	0	-1	-5	0	-1	-1	-5	-4	-2
E	0	-2	1	3	-5	2	**4**	0	0	-2	-3	-1	-2	-5	-1	0	-1	-7	-4	-2
G	1	-3	0	0	-3	-2	0	**4**	-3	-3	-4	-2	-3	-4	-1	1	-1	-7	-5	-2
H	-2	1	2	0	-3	2	0	-3	**6**	-3	-2	-1	-3	-2	-1	-1	-2	-3	0	-2
I	-1	-2	-2	-3	-2	-2	-2	-3	-3	**5**	2	-2	2	0	-2	-2	0	-5	-2	3
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	**5**	-3	3	1	-3	-3	-2	-2	-2	1
K	-2	3	1	0	-5	0	-1	-2	-1	-2	-3	**4**	0	-5	-2	-1	0	-4	-4	-3
M	-1	-1	-2	-3	-5	-1	-2	-3	-3	2	3	0	**7**	0	-2	-2	-1	-4	-3	1
F	-3	-4	-3	-6	-5	-5	-5	-4	-2	0	1	-5	0	**7**	-4	-3	-3	-1	5	-2
P	1	-1	-1	-2	-3	0	-1	-1	-1	-2	-3	-2	-2	-4	**5**	1	0	-5	-5	-2
S	1	-1	1	0	0	-1	0	1	-1	-2	-3	-1	-2	-3	1	**2**	1	-2	-3	-1
T	1	-1	0	-1	-2	-1	-1	-1	-2	0	-2	0	-1	-3	0	1	**3**	-5	-3	0
W	-5	1	-4	-6	-7	-5	-7	-7	-3	-5	-2	-4	-4	-1	-5	-2	-5	**12**	-1	-6
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-2	-2	-4	-3	5	-5	-3	-3	-1	**8**	-3
V	0	-3	-2	-3	-2	-2	-2	-2	-2	3	1	-3	1	-2	-2	-1	0	-6	-3	**4**

BLOSUM Derivation

$$
\begin{aligned}
q_{ij} &= \frac{a_{ij}}{I\binom{n}{2}} \\
e_{ij} &= \begin{cases} p_i^2, i = j \\ 2p_i p_j, i \neq j \end{cases} \\
s_{ij} &= log_2\left(q_{ij}/e_{ij}\right) \\
\iota_{ij} &= \text{round}\left(2s_{ij}\right)
\end{aligned}
$$

Entry in BLOCKS Database

```
ID   CLAUDIN1; BLOCK
AC   IPB003548C; distance from previous block=(3,4)
DE   Claudin-1 signature
BL   PR01377;  width=8; seqs=3; 99.5%=501; strength=972
CLD1_HUMAN|O95832  ( 152) MTPVNARY  83
CLD1_MOUSE|O88551  ( 152) LTPINARY 100

CLD1_RAT|P56745    ( 152) MTPVNARY  83
//
```

BLOSUM Example Computation: e_{ij}

A

i	Count	p_i
A	3	1/8
R	3	1/8
N	3	1/8
I	1	1/24
L	1	1/24
M	2	1/12
P	3	1/8
T	3	1/8
Y	3	1/8
V	2	1/12

B

	A	R	N	I	L	M	P	T	Y	V
A	1/64									
R	0	1/64								
N	0	0	1/64							
I	0	0	0	1/576						
L	0	0	0	0	1/576					
M	0	0	0	0	1/144	1/144				
P	0	0	0	0	0	0	1/64			
T	0	0	0	0	0	0	0	1/64		
Y	0	0	0	0	0	0	0	0	1/64	
V	0	0	0	1/144	0	0	0	0	0	1/144

BLOSUM Example Computation: q_{ij}

A

	A	R	N	I	L	M	P	T	Y	V
A	3									
R	0	3								
N	0	0	3							
I	0	0	0	0						
L	0	0	0	0	0					
M	0	0	0	0	2	1				
P	0	0	0	0	0	0	3			
T	0	0	0	0	0	0	0	3		
Y	0	0	0	0	0	0	0	0	3	
V	0	0	0	2	0	0	0	0	0	1

B

	A	R	N	I	L	M	P	T	Y	V
A	1/8									
R	0	1/8								
N	0	0	1/8							
I	0	0	0	0						
L	0	0	0	0	0					
M	0	0	0	0	1/8	1/24				
P	0	0	0	0	0	0	1/8			
T	0	0	0	0	0	0	0	1/8		
Y	0	0	0	0	0	0	0	0	1/8	
V	0	0	0	1/8	0	0	0	0	0	1/24

ICB © 2006 Birkhäuser Verlag

BLOSUM Example Computation: s_{ij} & ι_{ij}

A

	A	R	N	I	L	M	P	T	Y	V
A	3.0									
R	-	3.0								
N	-	-	3.0							
I	-	-	-	-						
L	-	-	-	-	-					
M	-	-	-	-	3.6	2.6				
P	-	-	-	-	-	-	3.0			
T	-	-	-	-	-	-	-	3.0		
Y	-	-	-	-	-	-	-	-	3.0	
V	-	-	-	3.6	-	-	-	-	-	2.6

B

	A	R	N	I	L	M	P	T	Y	V
A	6									
R	-	6								
N	-	-	6							
I	-	-	-	-						
L	-	-	-	-	-					
M	-	-	-	-	7	5				
P	-	-	-	-	-	-	3			
T	-	-	-	-	-	-	-	3		
Y	-	-	-	-	-	-	-	-	3	
V	-	-	-	7	-	-	-	-	-	5

BLOSUM Example Computation: Clustering

A

i	Count	p_i
A	3	1/8
R	3	1/8
N	3	1/8
I	1	1/32
L	1	1/32
M	2	3/32
P	3	1/8
T	3	1/8
Y	3	1/8
V	2	3/32

B

	A	R	N	I	L	M	P	T	Y	V
A	2									
R	0	2								
N	0	0	2							
I	0	0	0	0						
L	0	0	0	0	0					
M	0	0	0	0	1	1				
P	0	0	0	0	0	0	2			
T	0	0	0	0	0	0	0	2		
Y	0	0	0	0	0	0	0	0	2	
V	0	0	0	1	0	0	0	0	0	1

BLOSUM62 Substitution Matrix

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

Application BLOSUM80

Homeodomain

```
D   NSGRTNFTNKQLTELEKEFHFNRYLTRARR   537
H   NAVRTNFTTKQLTELEKEFHFNKYLTRARR   259
```

Homeodomain

```
D   IEIANTLQLNETQVKIWFQNRRMKQKKRVK   567
H   VEIAASLQLNETQVKIWFQNRRMKQKKREK   289
```

```
D   EGLIP   572
H   EGLLP   294
```

Application BLOSUM62

Homeodomain

```
D   NSGRTNFTNKQLTELEKEFHFNRYLTRARR   537
H   NAVRTNFTTKQLTELEKEFHFNKYLTRARR   259
```

Homeodomain

```
D   IEIANTLQLNETQVKIWFQNRRMKQKKRVK   567
H   VEIAASLQLNETQVKIWFQNRRMKQKKREK   289
```

```
D   EGLIPADILTQHSTSVISEKPPQQQQPQP   596
H   EGLLPISPATPPGNDEKAEESSEKSSSSP   318
```

Application BLOSUM45

```
D   SAAEYGITTSTTTGNPGTPLHPSSHSPADSYYESDSVHSYYATAAVATVAPPSNSSPITAANASA
H   SFLEYPILSSGDSGTCSARAYPSDHRI-----------TTFQSCAVSANSCGGDDRFLVGRGVQI

D   TS------NTQQQQQQAAIISSEN-GMMYTNLDC--MYPTAQAQAPVHGYAGQIEEK----YAAV
H   GSPHHHHHHHHHHPQPATYQTSGNLGVSYSHSSCGPSYGSQNFSAPYSPYALNQEADVSGGYPQC

D   LHASYAPGMVLEDQDPMMQQATQSQMWHHQQHLAGSYALDAMDSLGMHAHMHHGLPHGHLGNLAN
H   APAVYSGNL----SSPMVQ--------HH--------------------HHHQGYAGGAVGS---

D   NPHQQQPQVQQQQQQPHQQPQHPQNQSPAAHQQHHQNSVSPNGGMNRQQRGGVISPGSSTSSSTS
H   ------PQYIH-----HSYGQEHQSLALAT----YNNSLSP---LHASHQEACRSPASETSS---

D   ASNGAHPASTQSKSPNHSSSIPTYKWMQLKRNVPKPQAPSYLPAPKLPASGIASMHDYQMNGQLD
H   ------PAQ-------------TFDWMKVKRNPPK------------------------------

D   MCRGGGGGGSGVGNGPVG-VGGNGSPGIGGVLSVQNSLIMANSAAAGSAHPNGMGVGLGSGSGL
H   -------------TGKVGEYGYLGQP--------------------------------------
```

Homeodomain

```
D   SSCSLSSNTNNSGRTNFTNKQLTELEKEFHFNRYLTRARRIEIANTLQLNETQVKIWFQNRRMKQ
H   ----------NAVRTNFTTKQLTELEKEFHFNKYLTRARRVEIAASLQLNETQVKIWFQNRRMKQ
```

Homeodomain

Corresponding PAM & BLOSUM Matrices

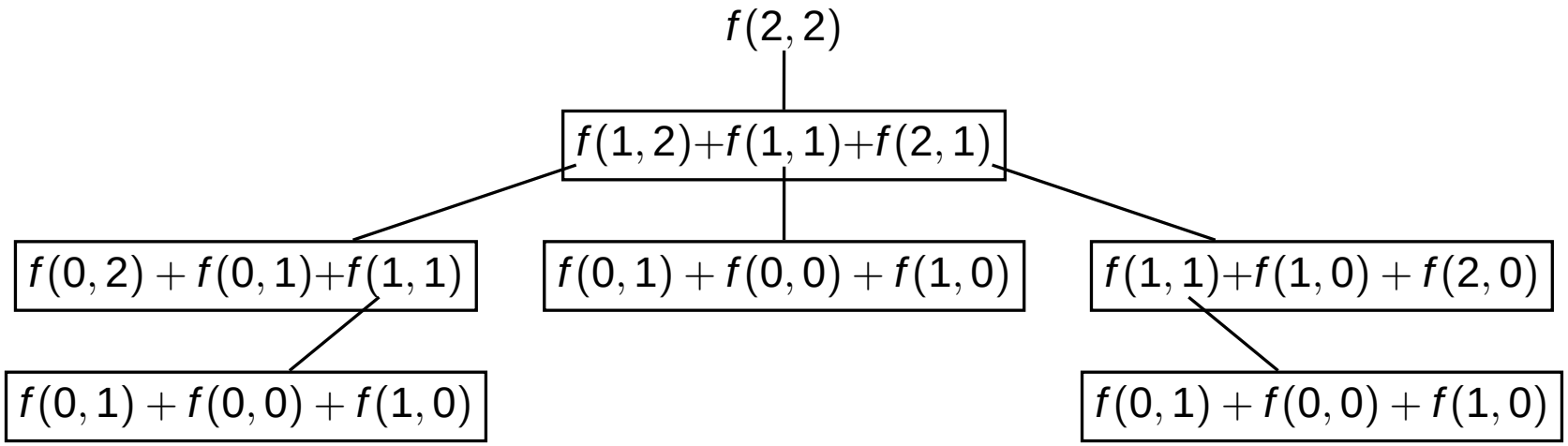
BLOSUM*n*	PAM*n*
45	250
62	160
80	120

ICB © 2006 Birkhäuser Verlag

Ends of Alignments

1. Overhang S_1: $\binom{S_1[m]}{-}$
2. Overhang S_2: $\binom{-}{S_2[n]}$
3. Flush: $\binom{S_1[m]}{S_2[n]}$

Number of Possible Global Alignments: Top-Down



Number of Posible Global Alignments: Bottom-Up

A

	0	1	2
0			
1			
2			

$\longrightarrow$

B

	0	1	2
0	**1**	**1**	1
1	**1**		
2	1		

$\longrightarrow$

C

	0	1	2
0	1	1	1
1	1	3	5
2	1	5	13

Notation

1. $|S|$ is the *length* of S.
2. $S[i..j]$ is a *substring* of S.
3. $S[1..i]$ is a *prefix* of S.
4. $S[i..|S|]$ is a *suffix* of S.

ICB © 2006 Birkhäuser Verlag

Global Alignment: Initialization

S_1

		–	A	C	C	G	T	T
		0	1	2	3	4	5	6
–	0	0	-1	-2	-3	-4	-5	-6
A	1	-1						
G	2	-2						
T	3	-3						
T	4	-4						
C	5	-5						
A	6	-6						

S_2

index of S_1

first row of $F(i,j)$

index of S_2

first column of $F(i,j)$

Global Alignment: Filling In

$F(i,j)$			A	C	C	G	T	T
		0	1	2	3	4	5	6
	0	0 ↘	-1 ↓	-2	-3	-4	-5	-6
A	1	-1 →						
G	2	-2						
T	3	-3						
T	4	-4						
C	5	-5						
A	6	-6						

Global Alignment: Filled In

$F(i,j)$			A	C	C	G	T	T
		0	1	2	3	4	5	6
	0	0	-1	-2	-3	-4	-5	-6
A	1	-1	1	0	-1	-2	-3	-4
G	2	-2	0	0	-1	0	-1	-2
T	3	-3	-1	-1	-1	-1	1	0
T	4	-4	-2	-2	-2	-2	0	2
C	5	-5	-3	-1	-1	-2	-1	1
A	6	-6	-4	-2	-2	-2	-2	**0**

ICB © 2006 Birkhäuser Verlag

Global Alignment: Recursion

$$
\begin{aligned}
F(i,0) &= i \times g_{\mathrm{e}}, 0 \leq i \leq m, \\
F(0,j) &= j \times g_{\mathrm{e}}, 0 \leq j \leq n, \\
F(i,j) &= \max \begin{cases} F(i,j-1) + g_{\mathrm{e}} \\ F(i-1,j-1) + s(S_1[i], S_2[j]) \\ F(i-1,j) + g_{\mathrm{e}} \end{cases}
\end{aligned}
$$

Global Alignment: Trace-back

		–	A	C	C	G	T	T
		0	1	2	3	4	5	6
–	0	**0**	← -1	← -2	← -3	← -4	← -5	← -6
A	1	↑ -1	↖ **1**	← **0**	← **-1**	← -2	← -3	← -4
G	2	↑ -2	↑ 0	↖ 0	↖ ← -1	↖ **0**	← -1	← -2
T	3	↑ -3	↑ -1	↖ ↑ -1	↖ -1	↑ -1	↖ **1**	↖ ← 0
T	4	↑ -4	↑ -2	↖ ↑ -2	↖ ↑ -2	↖ ↑ -2	↖ ↑ 0	↖ **2**
C	5	↑ -5	↖ ↑ -3	↑ -1	↑ -1	← -2	↑ -1	↑ **1**
A	6	↑ -6	↑ -4	↖ -2	↖ -2	← -2	↑ -2	↑ **0**

```
ACCGTT--
A--GTTCA
```

Genome Projects

Year	Organism	Genome Size	Commentary
1978	bacteriophage ϕX174	5386 bp	first genome [8]
1982	bacteriophage λ	48.502 bp	shotgun sequencing [9]
1995	*Haemophilus influenzae*	1.8 Mb	first free-living organism [5]
1996	*Saccharomyces cerevisiae*	12 Mb	first eucaryote [6]
1998	*Caenorhabditis elegans*	97 Mb	first metazoan [4]
2000	*Arabidopsis thaliana*	120 Mb	first plant [11]
2000	*Drosophila melanogaster*	180 Mb	first metazoan with whole-genome shotgun [1]
2001/2003	*Homo sapiens*	3.1 Gb	aim of the human genome project [7, 12, 2]
2002	*Mus musculus*	2.5 Gb	comparative pharmacogenomics [3]
2005	*Pan troglodytes*	3.1 Gb	closest relative of humans [10]

ICB © 2006 Birkhäuser Verlag

Shotgun Sequencing

L length of DNA sequence

s number of nucleotides sequenced

Probability of not sequencing a specific nucleotide

$$P_0 = \left(1 - \frac{1}{L}\right)^s \approx e^{-\frac{s}{L}}$$

Probability of sequencing a specific nucleotide

$$P = 1 - e^{-c},$$

where $c = s/L$ is the *coverage*.

Overlap Alignment

		–	A	C	C	G	T	T
		0	1	2	3	4	5	6
–	0	**0**	← **0**	← **0**	← 0	← 0	← 0	← 0
A	1	↑ 0	↖ 1	← 0	↖←↑ **-1**	↖↑ -1	↖↑ -1	↖↑ -1
G	2	↑ 0	↑ 0	↖ 0	↖← -1	↖ **0**	← -1	↖←↑ -2
T	3	↑ 0	↖←↑ -1	↖↑ -1	↖ -1	↑ -1	↖ **1**	↖← 0
T	4	↑ 0	↖← -1	↖←↑ -2	↖↑ -2	↖↑ -2	↖↑ 0	↖ **2**
C	5	↑ 0	↖← -1	↖ 0	↖← -1	← -2	↑ -1	↑ 1
A	6	↑ 0	↖ 1	← 0	↖← -1	↖← -2	↑ -2	↑ 0

ACCGTT--
--AGTTCA

Overlap Alignment: Recursion

$$
\begin{aligned}
F(i,0) &= 0 \\
F(0,j) &= 0 \\
F(i,j) &= \max \begin{cases} F(i,j-1) + g_{\mathrm{e}} \\ F(i-1,j-1) + s(S_1[i], S_2[j]) \\ F(i-1,j) + g_{\mathrm{e}} \end{cases}
\end{aligned}
$$

Local Alignment

		–	A	C	C	G	T	T
		0	1	2	3	4	5	6
–	0	0	0	0	0	0	0	0
A	1	0	↖ 1	0	**0**	0	0	0
G	2	0	0	0	0	↖ **1**	0	0
T	3	0	0	0	0	0	↖ **2**	↖ ← 1
T	4	0	0	0	0	0	↖ ↑ 1	↖ **3**
C	5	0	0	↖ 1	↖ 1	0	0	↑ 2
A	6	0	↖ 1	0	0	0	0	↑ 1

GTT

GTT

Local Alignment: Recursion

$$
\begin{aligned}
F(i,0) &= 0 \\
F(0,j) &= 0 \\
F(i,j) &= \max \begin{cases} F(i,j-1) + g_{\mathrm{e}} \\ F(i-1,j-1) + s(S_1[i], S_2[j]) \\ F(i-1,j) + g_{\mathrm{e}} \\ 0 \end{cases}
\end{aligned}
$$

Affine Gap Cost

F

	–	A	T	C
–	0	$-\infty$	$-\infty$	$-\infty$
A	$-\infty$	1	-10	-12
C	$-\infty$	-10	-2	-5

V

	–	A	T	C
–	0	-7	-9	-11
A	$-\infty$	$-\infty$	-6	-8
C	$-\infty$	$-\infty$	-17	-9

H

	–	A	T	C
–	0	$-\infty$	$-\infty$	$-\infty$
A	-7	$-\infty$	$-\infty$	$-\infty$
C	-9	-6	-17	-19

ATC
A-C

Recursion for Affine Gaps

$$
\begin{aligned}
F(0,0) &= 0 \\
F(i,0) &= V(i,0) = -\infty; i > 0 \\
F(0,j) &= H(0,j) = -\infty; j > 0 \\
H(i,0) &= g_{\mathrm{o}} + i \times g_{\mathrm{e}} \\
V(0,j) &= g_{\mathrm{o}} + j \times g_{\mathrm{e}} \\
F(i,j) &= \max \begin{cases} F(i-1,j-1) + s(S_1[i], S_2[j]) \\ H(i-1,j-1) + g_{\mathrm{o}} + g_{\mathrm{e}} \\ V(i-1,j-1) + g_{\mathrm{o}} + g_{\mathrm{e}} \end{cases} \\
H(i,j) &= \max \begin{cases} F(i,j-1) + g_{\mathrm{o}} + g_{\mathrm{e}} \\ H(i,j-1) + g_{\mathrm{e}} \end{cases} \\
V(i,j) &= \max \begin{cases} F(i-1,j) + g_{\mathrm{o}} + g_{\mathrm{e}} \\ V(i-1,j) + g_{\mathrm{e}} \end{cases}
\end{aligned}
$$

Comparison

Global, score = -52

local

```
D.m.  ACTTCACCAGCTCCCTGGCGGTAAGTTGAT---CAAA---GGAAACGCAAAGTTT   49
D.a.  GTTTCACTACTTCCTTTCGGGTAAGTAAATATATAAATATATAAAAATATAATTT   55

D.m.  TCAAG    54
D.a.  TCATC    60
```

Local, score = 7

```
D.m.  GGTAAGT     7
D.a.  GGTAAGT     7
```

Overlap, score = 1

```
D.m.  ACTTCACCAGCTCCCTGGCGGTAAGTTGATCAAAGGAAACGCAAAGTTTTCAAG-   54
D.a.  ---------------------------------------------------------GT    2

D.m.  ------------------------------------------------------------   54
D.a.  TTCACTACTTCCTTTCGGGTAAGTAAATATATAAATATATAAAAATATAATTTTC   57

D.m.  ---    54
D.a.  ATC    60
```

References

- M. D. Adams et al.
  The genome sequence of drosophila melanogaster.
  Science, 287:2185–2195, 2000.
- F. S. Collins, E. D. Green, A. E. Guttmacher, and M. S. Guyer.
  A blueprint for the genomic era.
  Nature, 422:835–847, 2003.
- Mouse Genome Sequencing Consortium.
  Initial sequencing and comparative analysis of the mouse genome.
  Nature, 420:520–561, 2002.
- The C. elegans Sequencing Consortium.
  Genome sequence of the nematode *C. elegans*: a platform for investigating biology.
  Science, 282:2012–2018, 1998.
- R. D. Fleischmann et al.
  Whole-genome random sequencing and assembly of haemophilus influenzae rd.
  Science, 269:496–512, 1995.
- A. Goffeau and et al.
  Life with 6000 genes.
  Science, 274:546–567, 1996.
- International Human Genome Sequencing Consortium.
  Initial sequencing and analysis of the human genome.
  Nature, 409:860–921, 2001.
- F. Sanger, A. R. Coulson, T. Friedmann, G. M. Air, B. G. Barrell, N. L. Brown, J. C. Fiddes, C. A. Hutchison III, P. M. Slocombe, and M. Smith.
  The nucleotide sequence of bacteriphage fx174.
  Journal of Molecular Biology, 125:225–246, 1978.
- F. Sanger, A. R. Coulson, G. F. Hong, D. F. Hill, and G. B. Petersen.
  Nucleotide sequence of bacteriphage l DNA.
  Journal of Molecular Biology, 162:729–773, 1982.
- The Chimpanzee Sequencing and Analysis Consortium.
  Initial sequence of the chimpanzee genome and comparison with the human genome.
  Nature, 437:69–87, 2005.
- The Arabidopsis Genome Initiative.
  Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*.
  Nature, 408:796–815, 2000.
- J.C. Venter et al.
  The sequence of the human genome.
  Science, 291:1304–1351, 2001.

ICB © 2006 Birkhäuser Verlag