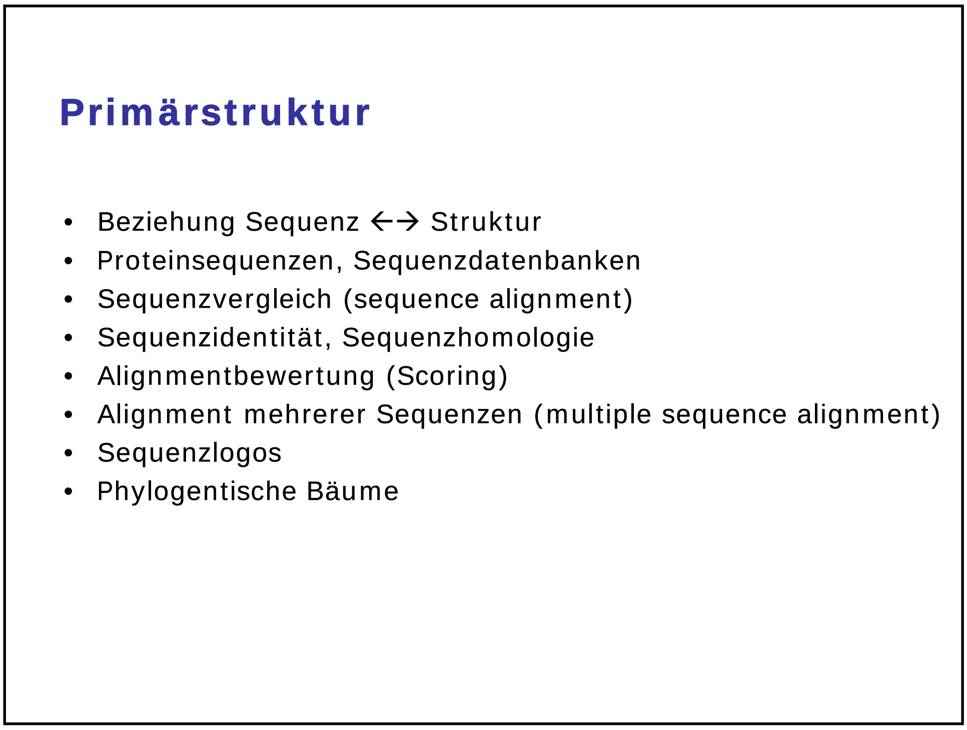
Primärstruktur

Wintersemester 2011/12

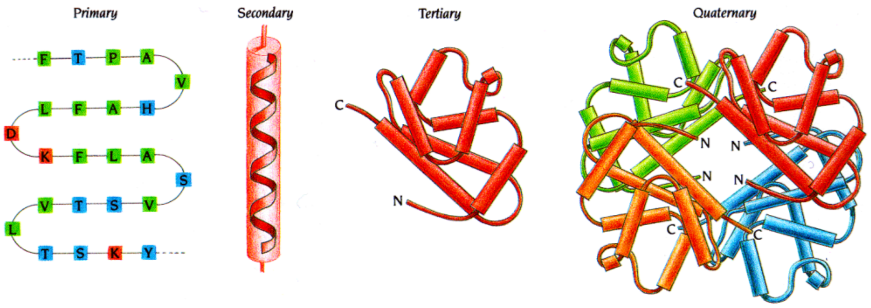
Peter Güntert



Primärstruktur

- Beziehung Sequenz <-> Struktur
- Proteinsequenzen, Sequenzdatenbanken
- Sequenzvergleich (sequence alignment)
- Sequenzidentität, Sequenzhomologie
- Alignmentbewertung (Scoring)
- Alignment mehrerer Sequenzen (multiple sequence alignment)
- Sequenzlogos
- Phylogentische Bäume

Sequenz → Struktur



- Die Sequenz bestimmt die dreidimensionale Struktur.
- Proteine mit ähnlicher Sequenz haben ähnliche Struktur.
- Aber: Auch Proteine mit unterschiedlicher Sequenz können ähnliche Strukturen haben.
- Proteinstrukturen sind evolutionär besser konserviert als Sequenzen.

Sequence identity → Structural similarity

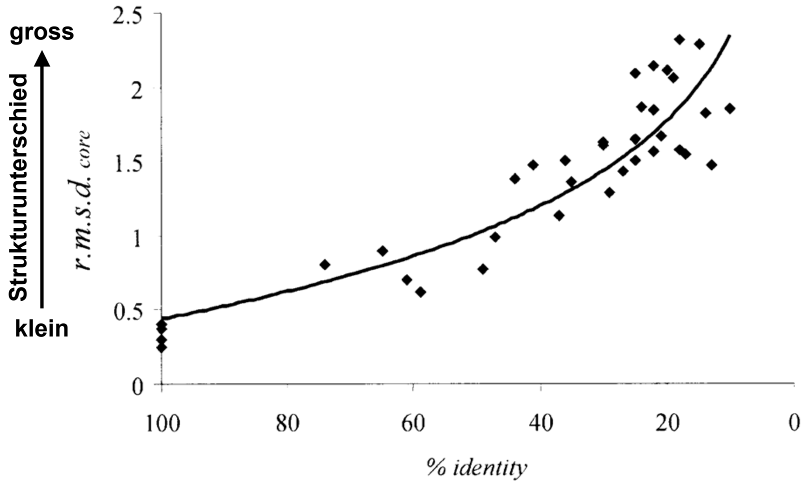


Figure 1.23 Relationship between sequence identity and structural similarity. The plot is obtained using the same set of proteins originally analyzed by Lesk and Chothia.

Sequence identity → Structural similarity

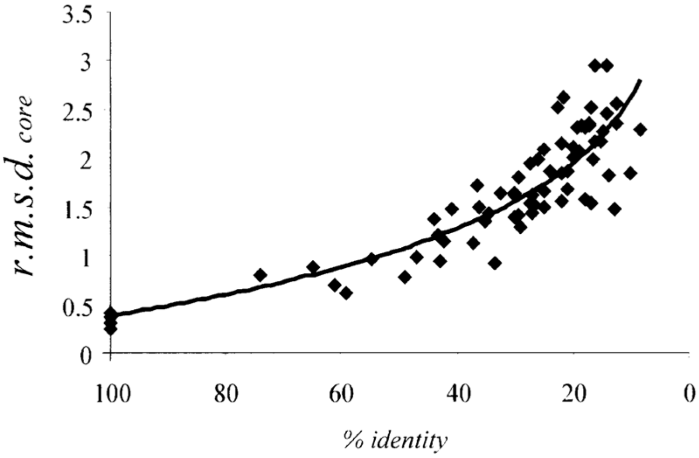
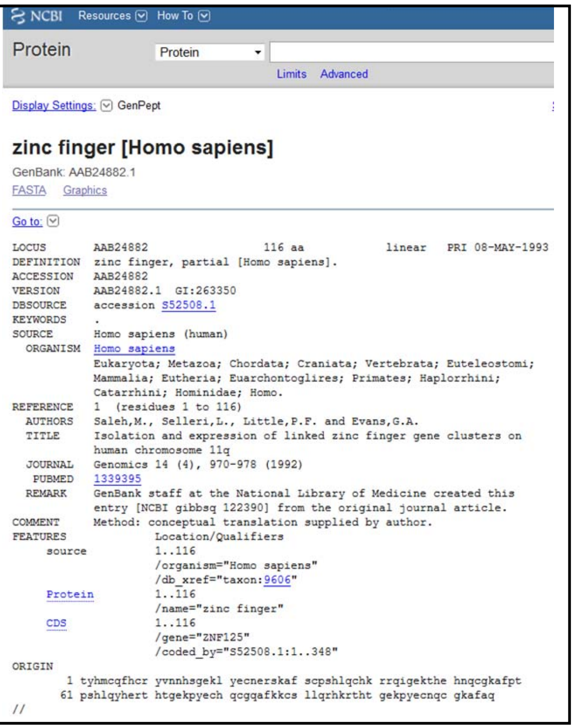


Figure 1.25 Relationships between sequence identity and structural similarity. The plot was obtained by using a larger set of proteins than in Figure 1.23, but the trend is essentially the same.

Sequenz-datenbank

www.ncbi.nlm.nih.gov/protein

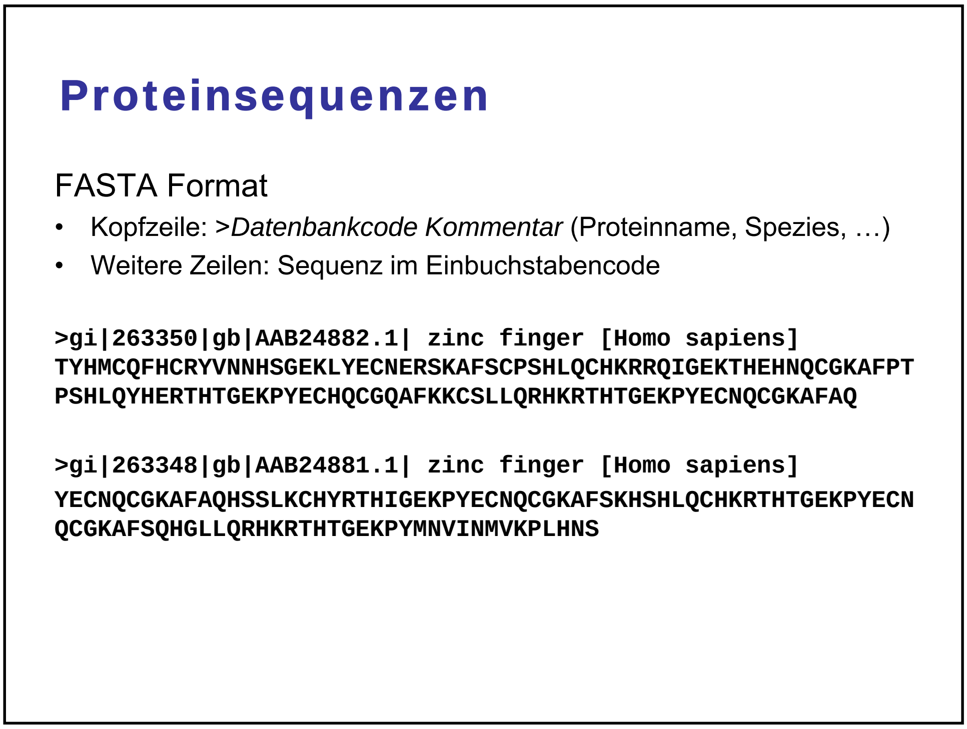
“The Protein database is a collection of sequences from several sources, including translations from annotated coding regions in GenBank, RefSeq and TPA, as well as records from SwissProt, PIR, PRF, and PDB. Protein sequences are the fundamental determinants of biological structure and function.”



zinc finger [Homo sapiens]

GenBank: AAB24882.1

```
LOCUS       AAB24882                 116 aa            linear   PRI 08-MAY-1993
DEFINITION  zinc finger, partial [Homo sapiens].
ACCESSION   AAB24882
VERSION     AAB24882.1  GI:263350
DBSOURCE    accession S52508.1
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 116)
  AUTHORS   Saleh,M., Selleri,L., Little,P.F. and Evans,G.A.
  TITLE     Isolation and expression of linked zinc finger gene clusters on
            human chromosome 11q
  JOURNAL   Genomics 14 (4), 970-978 (1992)
   PUBMED   1339395
  REMARK    GenBank staff at the National Library of Medicine created this
            entry [NCBI gibbsq 122390] from the original journal article.
COMMENT     Method: conceptual translation supplied by author.
FEATURES             Location/Qualifiers
     source          1..116
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
     Protein         1..116
                     /name="zinc finger"
     CDS             1..116
                     /gene="ZNF125"
                     /coded_by="S52508.1:1..348"
ORIGIN
        1 tyhmcqfhcr yvnnhsgekl yecnerskaf scpshlqchk rrqigekthe hnqcgkafpt
       61 pshlqyhert htgekpyech qcgqafkkcs llqrhkrtht gekpyecnqc gkafaq
//
```

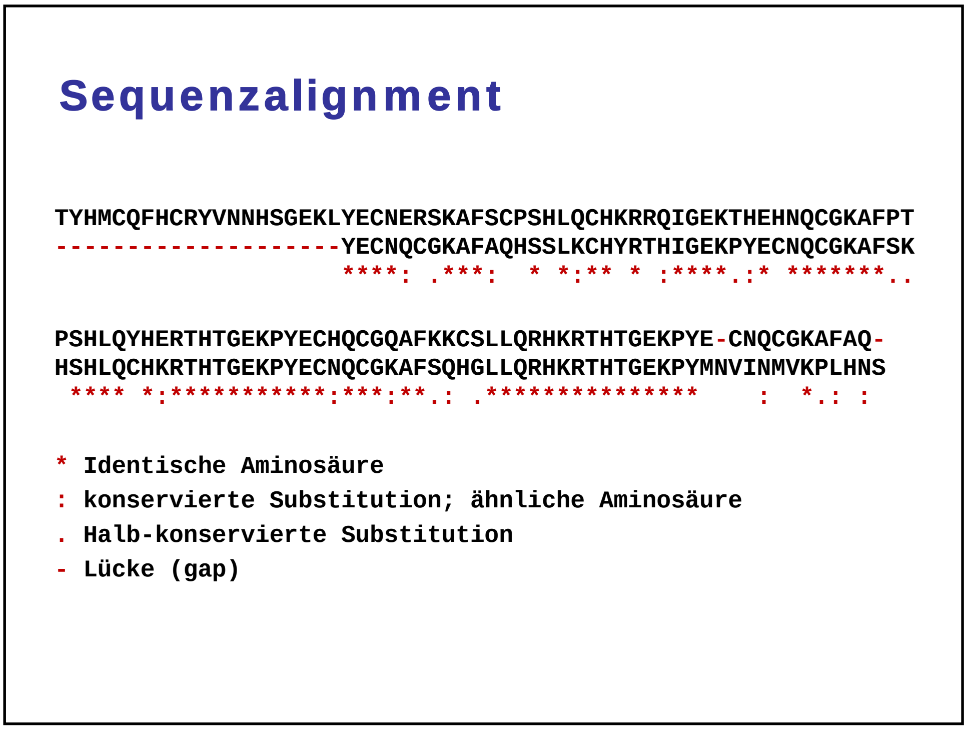


Proteinsequenzen

FASTA Format

- Kopfzeile: >*Datenbankcode Kommentar* (Proteinname, Spezies, …)
- Weitere Zeilen: Sequenz im Einbuchstabencode

```
>gi|263350|gb|AAB24882.1| zinc finger [Homo sapiens]
TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT
PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYECNQCGKAFAQ

>gi|263348|gb|AAB24881.1| zinc finger [Homo sapiens]
YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSKHSHLQCHKRTHTGEKPYECN
QCGKAFSQHGLLQRHKRTHTGEKPYMNVINMVKPLHNS
```



Sequenzalignment

```
TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT
--------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK
                    ****: .***:  * *:** * :****.:* *******..

PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE-CNQCGKAFAQ-
HSHLQCHKRTHTGEKPYECNQCGKAFSQHGLLQRHKRTHTGEKPYMNVINMVKPLHNS
 **** *:***********:***:**.: .***************    :  *.: :

* Identische Aminosäure
: konservierte Substitution; ähnliche Aminosäure
. Halb-konservierte Substitution
- Lücke (gap)
```

Sequenzidentität

- Definition:

$$Sequenzidentität = \frac{Anzahl\ identischer\ AS}{\min\limits_{Sequenzen}(Anzahl\ AS)} \times 100\%$$

- Beispiel:

```
TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT
--------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK
                    ****: .***:  * *:** * :****.:* *******..

PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE-CNQCGKAFAQ-
HSHLQCHKRTHTGEKPYECNQCGKAFSQHGLLQRHKRTHTGEKPYMNVINMVKPLHNS
 **** *:***********:***:**.: .***************    :  *.: :
```

Anzahl identischer AS: 61

Länge der Sequenzen: 116 AS, 98 AS

→ **Sequenzidentität = 60/98 × 100% = 62,2%**

Sequenzhomologie

- Definition:

$$Sequenzhomologie = \frac{Anzahl\ homologer\ AS}{\min\limits_{Sequenzen}(Anzahl\ AS)} \times 100\%$$

- Welche AS sind homolog (aehnlich)?
- Beispiel:

```
TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT
--------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK
                    ****: .***:  * *:** * :****.:* *******..
PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE-CNQCGKAFAQ-
HSHLQCHKRTHTGEKPYECNQCGKAFSQHGLLQRHKRTHTGEKPYMNVINMVKPLHNS
 **** *:***********:***:**.: .***************    :  *.: :
```

Anzahl homologer AS (* und :) : 61 + 12 = 73

Länge der Sequenzen: 116 AS, 98 AS

→ **Sequenzhomologie = 73/98 × 100% = 74,5%**

Sequenzidentität zufälliger Sequenzen

- Annahme: Alle 20 AS kommen gleich häufig mit Wahrscheinlichkeit p = 1/20 vor.

  → Erwartete Sequenzidentität fuer zwei gleich lange zufällige Sequenzen = p = 5%
- In natürlichen Proteinen kommen die AS mit unterschiedlichen Häufigkeiten p_1, ..., p_{20} vor (in %):

A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
8.3	5.7	4.4	5.3	1.7	4.0	6.2	7.2	2.2	5.2	9.0	5.7	2.4	3.9	5.1	6.9	5.8	1.3	3.2	6.6

→ Erwartete Sequenzidentität fuer zwei gleich lange zufällige Sequenzen

$$\text{Sequenzidentität} = \sum_{i=1}^{20} {p_i}^2 \approx 5.87\%$$

Globales und lokales Sequenzalignment

- Globales Sequenzalignment:
  - Optimales Alignment der gesamten Sequenzen
  - Gut fuer relativ ähnliche und ähnlich lange Sequenzen

```
--T--CC-C-AGT—-TATGT-CAGGGGACACG—A-GCATGCAGA-GAC
  |  || |  ||  | | | |||    ||  | | | | ||||   |
AATTGCCGCC-GTCGT-T-TTCAG----CA-GTTATG—T-CAGAT—C
```

- Lokales Sequenzalignment:
  - Optimales Alignment fuer Teilsequenz(en)
  - Gut zum Finden ähnlicher Teilsequenzen in längeren, unterschiedlichen Sequenzen

```
                  tccCAGTTATGTCAGgggacacgagcatgcagagac
                     ||||||||||||
aattgccgccgtcgttttcagCAGTTATGTCAGatc
```

Alignmentbewertung (Scoring)

- Einfaches Schema:
  - Identische AS (match): +1
  - Unterschiedliche AS (mismatch): -µ
  - Insertionen/Deletionen (indel): -σ

  →Score = #matches – µ × #mismatches – σ × #indels

- Verallgemeinerung:
  Scoringmatrix *S*(*i*,*j*) mit 21 × 21 Elementen (20 AS + indel)

	A	R	N	K
A	5	-2	-1	-1
R	-	7	-1	3
N	-	-	7	0
K	-	-	-	6

Beispiel (eines Teils) einer Scoringmatrix:
- Diagonalelemente gross
- Nichtdiagonalelemente meist negativ
- Austausch ähnlicher AS positiv (z.B. R → K)

Scoringmatrix

Blosum50

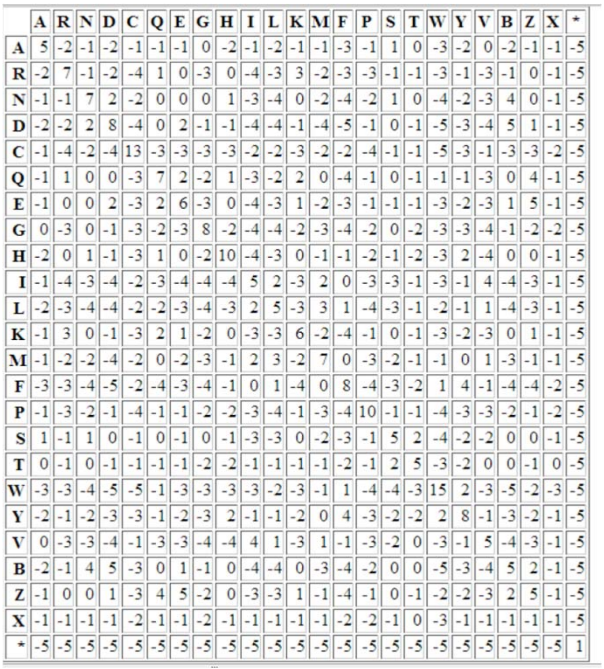
Scoringmatrizen können aus den Häufigkeiten für AS-Substitutionen in verwandten Sequenzen abgeleitet werden.

Log-odds score:

$$S(i,j) = \log \frac{P(i \to j)}{p_j}$$

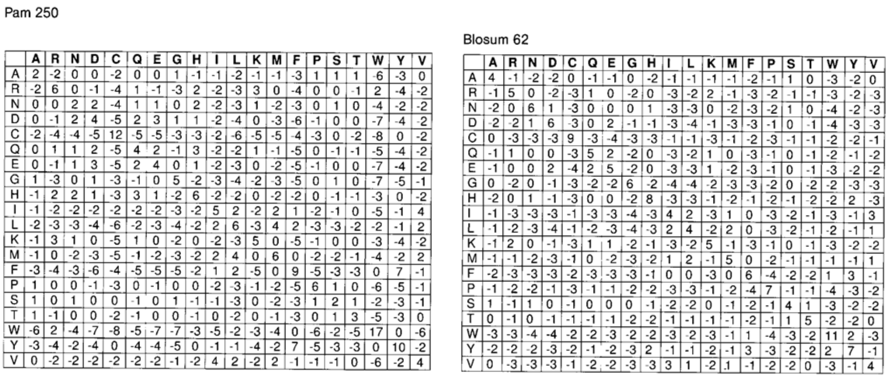
$P(i \to j)$: Wahrscheinlichkeit der Substitution (Mutation) von AS *i* zu *j*

p_j: Häufigkeit der AS j



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	5	-2	-1	-2	-1	-1	-1	0	-2	-1	-2	-1	-1	-3	-1	1	0	-3	-2	0	-2	-1	-1	-5
R	-2	7	-1	-2	-4	1	0	-3	0	-4	-3	3	-2	-3	-3	-1	-1	-3	-1	-3	-1	0	-1	-5
N	-1	-1	7	2	-2	0	0	0	1	-3	-4	0	-2	-4	-2	1	0	-4	-2	-3	4	0	-1	-5
D	-2	-2	2	8	-4	0	2	-1	-1	-4	-4	-1	-4	-5	-1	0	-1	-5	-3	-4	5	1	-1	-5
C	-1	-4	-2	-4	13	-3	-3	-3	-3	-2	-2	-3	-2	-2	-4	-1	-1	-5	-3	-1	-3	-3	-2	-5
Q	-1	1	0	0	-3	7	2	-2	1	-3	-2	2	0	-4	-1	0	-1	-1	-1	-3	0	4	-1	-5
E	-1	0	0	2	-3	2	6	-3	0	-4	-3	1	-2	-3	-1	-1	-1	-3	-2	-3	1	5	-1	-5
G	0	-3	0	-1	-3	-2	-3	8	-2	-4	-4	-2	-3	-4	-2	0	-2	-3	-3	-4	-1	-2	-2	-5
H	-2	0	1	-1	-3	1	0	-2	10	-4	-3	0	-1	-1	-2	-1	-2	-3	2	-4	0	0	-1	-5
I	-1	-4	-3	-4	-2	-3	-4	-4	-4	5	2	-3	2	0	-3	-3	-1	-3	-1	4	-4	-3	-1	-5
L	-2	-3	-4	-4	-2	-2	-3	-4	-3	2	5	-3	3	1	-4	-3	-1	-2	-1	1	-4	-3	-1	-5
K	-1	3	0	-1	-3	2	1	-2	0	-3	-3	6	-2	-4	-1	0	-1	-3	-2	-3	0	1	-1	-5
M	-1	-2	-2	-4	-2	0	-2	-3	-1	2	3	-2	7	0	-3	-2	-1	-1	0	1	-3	-1	-1	-5
F	-3	-3	-4	-5	-2	-4	-3	-4	-1	0	1	-4	0	8	-4	-3	-2	1	4	-1	-4	-4	-2	-5
P	-1	-3	-2	-1	-4	-1	-1	-2	-2	-3	-4	-1	-3	-4	10	-1	-1	-4	-3	-3	-2	-1	-2	-5
S	1	-1	1	0	-1	0	-1	0	-1	-3	-3	0	-2	-3	-1	5	2	-4	-2	-2	0	0	-1	-5
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	2	5	-3	-2	0	0	-1	0	-5
W	-3	-3	-4	-5	-5	-1	-3	-3	-3	-3	-2	-3	-1	1	-4	-4	-3	15	2	-3	-5	-2	-3	-5
Y	-2	-1	-2	-3	-3	-1	-2	-3	2	-1	-1	-2	0	4	-3	-2	-2	2	8	-1	-3	-2	-1	-5
V	0	-3	-3	-4	-1	-3	-3	-4	-4	4	1	-3	1	-1	-3	-2	0	-3	-1	5	-4	-3	-1	-5
B	-2	-1	4	5	-3	0	1	-1	0	-4	-4	0	-3	-4	-2	0	0	-5	-3	-4	5	2	-1	-5
Z	-1	0	0	1	-3	4	5	-2	0	-3	-3	1	-1	-4	-1	0	-1	-2	-2	-3	2	5	-1	-5
X	-1	-1	-1	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-2	-2	-1	0	-3	-1	-1	-1	-1	-1	-5
*	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	1

Amino acid substitution matrices



Pam 250

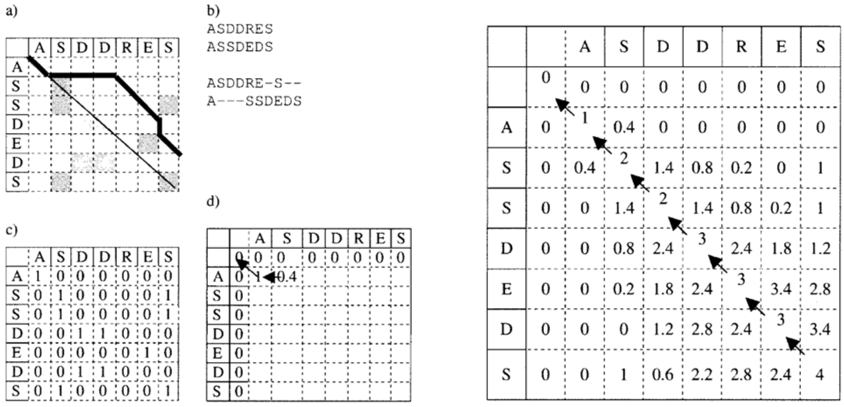
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2	-2	0	0	-2	0	0	1	-1	-1	-2	-1	-1	-3	1	1	1	-6	-3	0
R	-2	6	0	-1	-4	1	-1	-3	2	-2	-3	3	0	-4	0	0	-1	2	-4	-2
N	0	0	2	2	-4	1	1	0	2	-2	-3	1	-2	-3	0	1	0	-4	-2	-2
D	0	-1	2	4	-5	2	3	1	1	-2	-4	0	-3	-6	-1	0	0	-7	-4	-2
C	-2	-4	-4	-5	12	-5	-5	-3	-3	-2	-6	-5	-5	-4	-3	0	-2	-8	0	-2
Q	0	1	1	2	-5	4	2	-1	3	-2	-2	1	-1	-5	0	-1	-1	-5	-4	-2
E	0	-1	1	3	-5	2	4	0	1	-2	-3	0	-2	-5	-1	0	0	-7	-4	-2
G	1	-3	0	1	-3	-1	0	5	-2	-3	-4	-2	-3	-5	0	1	0	-7	-5	-1
H	-1	2	2	1	-3	3	1	-2	6	-2	-2	0	-2	-2	0	-1	-1	-3	0	-2
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5	2	-2	2	1	-2	-1	0	-5	-1	4
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6	-3	4	2	-3	-3	-2	-2	-1	2
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5	0	-5	-1	0	0	-3	-4	-2
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6	0	-2	-2	-1	-4	-2	2
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9	-5	-3	-3	0	7	-1
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6	1	0	-6	-5	-1
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2	1	-2	-3	-1
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3	-5	-3	0
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17	0	-6
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	-2
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4

Blosum 62

	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

Figure 4.5 The PAM250 (part *a*) and BLOSUM62 (part *b*) substitution matrices. The values corresponding to pairs of amino acids can be used to fill the alignment matrix (part *c* of Figure 4.4).

Needleman-Wunsch alignment algorithm



b)

```
ASDDRES
ASSDEDS

ASDDRE-S--
A---SSDEDS
```

c)

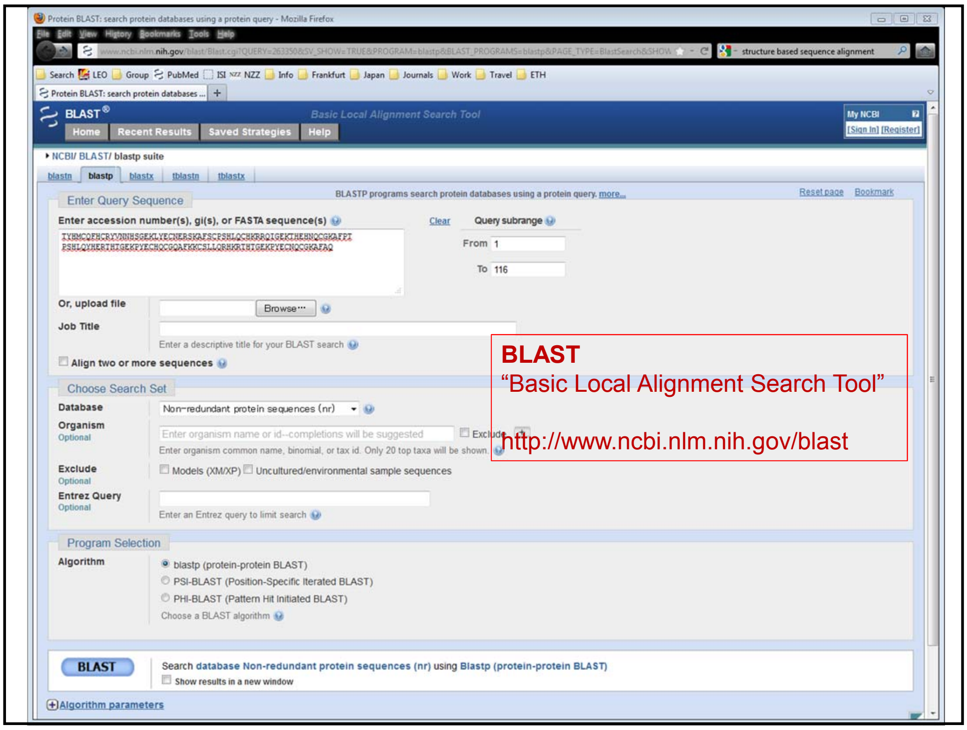
	A	S	D	D	R	E	S
A	1	0	0	0	0	0	0
S	0	1	0	0	0	0	1
S	0	1	0	0	0	0	1
D	0	0	1	1	0	0	0
E	0	0	0	0	0	1	0
D	0	0	1	1	0	0	0
S	0	1	0	0	0	0	1

d)

		A	S	D	D	R	E	S
	0	0	0	0	0	0	0	0
A	0	↖1	←0.4					
S	0							
S	0							
D	0							
E	0							
D	0							
S	0							

		A	S	D	D	R	E	S
	0	0	0	0	0	0	0	0
A	0	↖1	0.4	0	0	0	0	0
S	0	0.4	↖2	1.4	0.8	0.2	0	1
S	0	0	1.4	↖2	1.4	0.8	0.2	1
D	0	0	0.8	2.4	↖3	2.4	1.8	1.2
E	0	0	0.2	1.8	2.4	↖3	3.4	2.8
D	0	0	0	1.2	2.8	2.4	↖3	3.4
S	0	0	1	0.6	2.2	2.8	2.4	↖4

Figure 4.4 The Needleman and Wunsch alignment algorithm. A path in the matrix corresponds to an alignment. In the example, the thin line in part *a* of the figure corresponds to the first alignment shown in part *b*. The line runs diagonally and therefore corresponds to an alignment where there are no insertions or deletions. The tick line, instead, contains an horizontal line (indicating that the amino acids SDD of the first sequence do not correspond to any amino acid of the second and therefore represent an insertion in the first sequence) and two vertical lines (implying that the amino acid D and the final DS pair of the second sequence do not correspond to any amino acid in the first and is an insertion in the second sequence or, equivalently, a deletion in the first). To compute the optimum alignment we fill the cells of the matrix (part *c*) with a number representing the likelihood that the amino acid in the row is replaced by that in the column. In this example we assign 1 to identical amino acids and 0 to different ones. Part *d* shows the construction of the cumulative matrix as described in the text.


Protein BLAST: search protein databases using a protein query - Mozilla Firefox
File Edit View History Bookmarks Tools Help
www.ncbi.nlm.nih.gov/blast/Blast.cgi?QUERY=263350&SV_SHOW=TRUE&PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&SHOW
structure based sequence alignment
Search LEO Group PubMed ISI NZZ Info Frankfurt Japan Journals Work Travel ETH
Protein BLAST: search protein databases ...
BLAST®
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
[Sign In] [Register]
NCBI/ BLAST/ blastp suite
blastn
blastp
blastx
tblastn
tblastx
Reset page
Bookmark
Clear
Browse···
BLAST
Algorithm parameters

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

```
TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT
PSHLQYHERTHIGEKPYECHQCGQAFKKCSLLQRHKRIHTGEKPYECNQCGKAFAQ
```

Query subrange

From: 1

To: 116

Or, upload file

Job Title: Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism (Optional): Enter organism name or id--completions will be suggested □ Exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude (Optional): □ Models (XM/XP) □ Uncultured/environmental sample sequences

Entrez Query (Optional): Enter an Entrez query to limit search

Program Selection

Algorithm:
- ● blastp (protein-protein BLAST)
- ○ PSI-BLAST (Position-Specific Iterated BLAST)
- ○ PHI-BLAST (Pattern Hit Initiated BLAST)

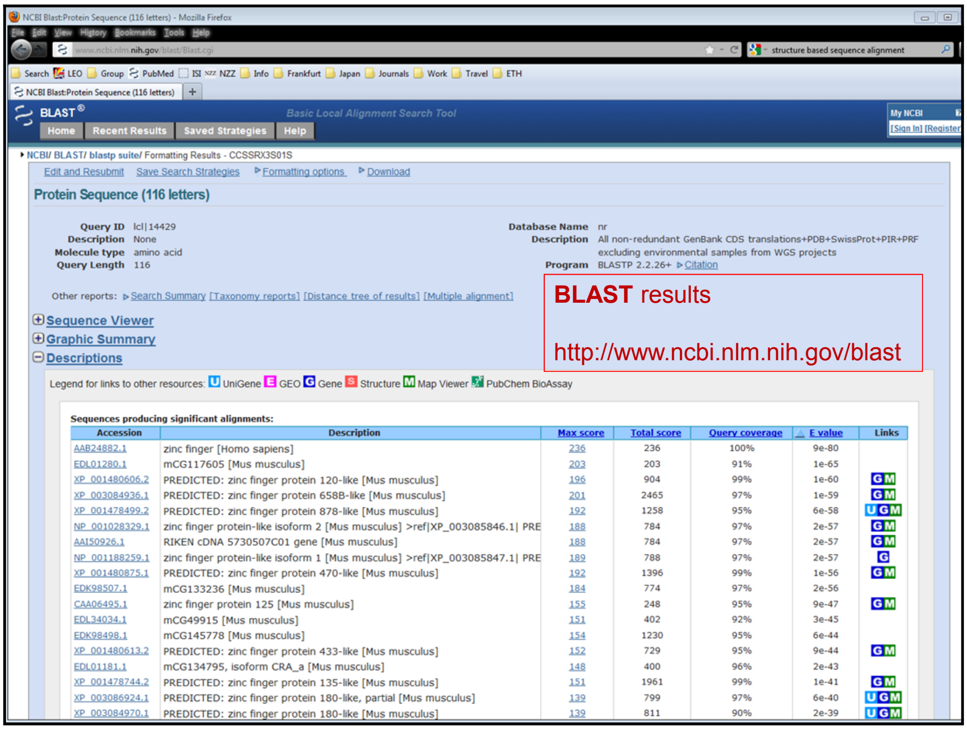
Choose a BLAST algorithm

Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)

□ Show results in a new window

BLAST
"Basic Local Alignment Search Tool"

http://www.ncbi.nlm.nih.gov/blast


NCBI Blast:Protein Sequence (116 letters) - Mozilla Firefox
File Edit View History Bookmarks Tools Help
www.ncbi.nlm.nih.gov/blast/Blast.cgi
structure based sequence alignment
Search LEO Group PubMed ISI NZZ Info Frankfurt Japan Journals Work Travel ETH
NCBI Blast:Protein Sequence (116 letters)
BLAST®
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
[Sign In] [Register]
NCBI/ BLAST/ blastp suite/ Formatting Results - CCSSRX3S01S
Edit and Resubmit
Save Search Strategies
Formatting options
Download

Protein Sequence (116 letters)

Query ID: lcl|14429
Description: None
Molecule type: amino acid
Query Length: 116

Database Name: nr
Description: All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Program: BLASTP 2.2.26+ Citation

Other reports: Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment]

⊞ Sequence Viewer
⊞ Graphic Summary
⊟ Descriptions

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer PubChem BioAssay

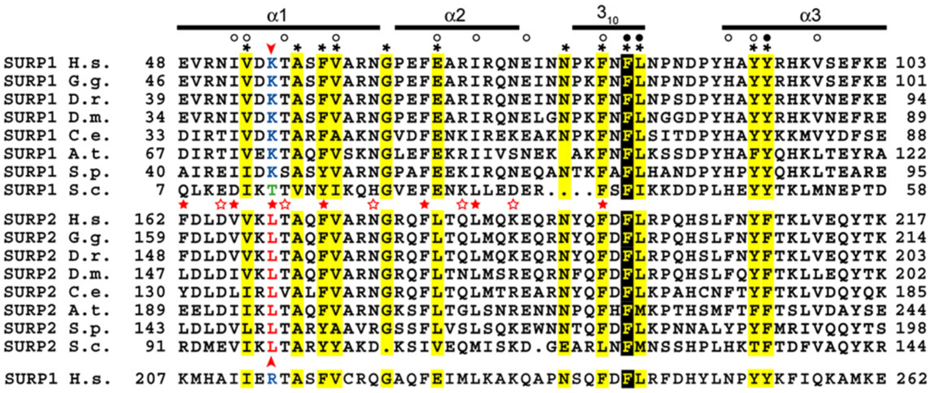
Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Links
AAB24882.1	zinc finger [Homo sapiens]	236	236	100%	9e-80	
EDL01280.1	mCG117605 [Mus musculus]	203	203	91%	1e-65	
XP_001480606.2	PREDICTED: zinc finger protein 120-like [Mus musculus]	196	904	99%	1e-60	G M
XP_003084936.1	PREDICTED: zinc finger protein 658B-like [Mus musculus]	201	2465	97%	1e-59	G M
XP_001478499.2	PREDICTED: zinc finger protein 878-like [Mus musculus]	192	1258	95%	6e-58	U G M
NP_001028329.1	zinc finger protein-like isoform 2 [Mus musculus] >ref|XP_003085846.1| PRE	188	784	97%	2e-57	G M
AAI50926.1	RIKEN cDNA 5730507C01 gene [Mus musculus]	188	784	97%	2e-57	G M
NP_001188259.1	zinc finger protein-like isoform 1 [Mus musculus] >ref|XP_003085847.1| PRE	189	788	97%	2e-57	G
XP_001480875.1	PREDICTED: zinc finger protein 470-like [Mus musculus]	192	1396	99%	1e-56	G M
EDK98507.1	mCG133236 [Mus musculus]	184	774	97%	2e-56	
CAA06495.1	zinc finger protein 125 [Mus musculus]	155	248	95%	9e-47	G M
EDL34034.1	mCG49915 [Mus musculus]	151	402	92%	3e-45	
EDK98498.1	mCG145778 [Mus musculus]	154	1230	95%	6e-44	
XP_001480613.2	PREDICTED: zinc finger protein 433-like [Mus musculus]	152	729	95%	9e-44	G M
EDL01181.1	mCG134795, isoform CRA_a [Mus musculus]	148	400	96%	2e-43	
XP_001478744.2	PREDICTED: zinc finger protein 135-like [Mus musculus]	151	1961	99%	1e-41	G M
XP_003086924.1	PREDICTED: zinc finger protein 180-like, partial [Mus musculus]	139	799	97%	6e-40	U G M
XP_003084970.1	PREDICTED: zinc finger protein 180-like [Mus musculus]	139	811	90%	2e-39	U G M

BLAST results

http://www.ncbi.nlm.nih.gov/blast

Alignment mehrerer Sequenzen

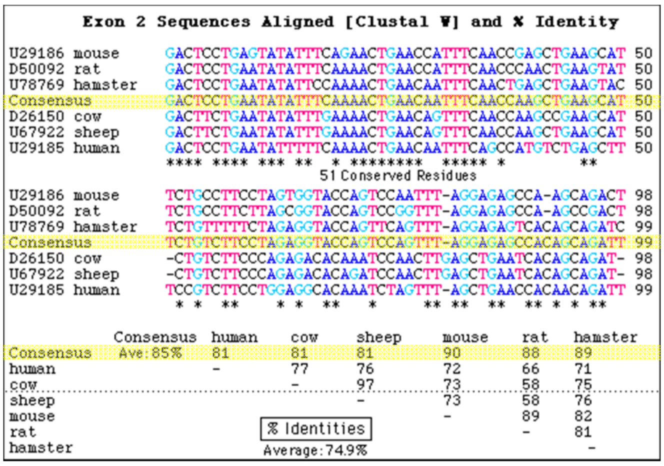


```
SURP1 H.s.   48 EVRNIVDKTASFVARNGPEFEARIRQNEINNPKFNFLNPNDPYHAYYRHKVSEFKE 103
SURP1 G.g.   46 EVRNIVDKTASFVARNGPEFEARIRQNEINNPKFNFLNPNDPYHAYYRHKVSEFKE 101
SURP1 D.r.   39 EVRNIVDKTASFVARNGPEFEARIRQNEINNPKFNFLNPSDPYHAYYRHKVNEFKE  94
SURP1 D.m.   34 EVRNIVDKTASFVARNGPEFEARIRQNELGNPKFNFLNGGDPYHAYYRHKVNEFRE  89
SURP1 C.e.   33 DIRTIVDKTARFAAKNGVDFENKIREKEAKNPKFNFLSITDPYHAYYKKMVYDFSE  88
SURP1 A.t.   67 DIRTIVEKTAQFVSKNGLEFEKRIIVSNEK AKFNFLKSSDPYHAFYQHKLTEYRA 122
SURP1 S.p.   40 AIREIIDKSASYVARNGPAFEEKIRQNEQANTKFAFLHANDPYHPYYQHKLTEARE  95
SURP1 S.c.    7 QLKEDIKTTVNYIKQHGVEFENKLLEDER....FSFIKKDDPLHEYYTKLMNEPTD  58

SURP2 H.s.  162 FDLDVVKLTAQFVARNGRQFLTQLMQKEQRNYQFDFLRPQHSLFNYFTKLVEQYTK 217
SURP2 G.g.  159 FDLDVVKLTAQFVARNGRQFLTQLMQKEQRNYQFDFLRPQHSLFNYFTKLVEQYTK 214
SURP2 D.r.  148 FDLDVVKLTAQFVARNGRQFLTQLMQKEQRNYQFDFLRPQHSLFNYFTKLVEQYTK 203
SURP2 D.m.  147 LDLDIVKLTAQFVARNGRQFLTNLMSREQRNFQFDFLRPQHSLFQYFTKLLEQYTK 202
SURP2 C.e.  130 YDLDLIRLVALFVARNGRQFLTQLMTREARNYQFDFLKPAHCNFTYFTKLVDQYQK 185
SURP2 A.t.  189 EELDIIKLTAQFVARNGKSFLTGLSNRENNNPQFHFMKPTHSMFTFFTSLVDAYSE 244
SURP2 S.p.  143 LDLDVLRLTARYAAVRGSSFLVSLSQKEWNNTQFDFLKPNNALYPYFMRIVQQYTS 198
SURP2 S.c.   91 RDMEVIKLTARYYAKD.KSIVEQMISKD.GEARLNFMNSSHPLHKTFTDFVAQYKR 144

SURP1 H.s.  207 KMHAIIERTASFVCRQGAQFEIMLKAKQAPNSQFDFLRFDHYLNPYYKFIQKAMKE 262
```

Konsensussequenz

Sequenz, die im Mittel am wenigsten von den verglichenen Einzelsequenzen abweicht.

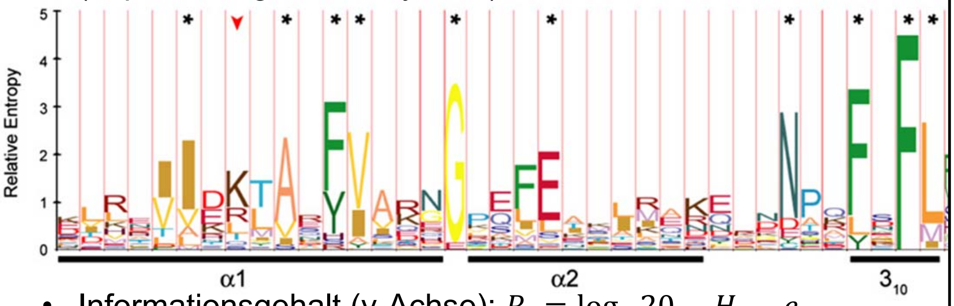
Beispiel: Ausschnitt aus dem Prionprotein PrP von Säugetieren.



```
        Exon 2 Sequences Aligned [Clustal W] and % Identity

U29186 mouse      GACTCCTGAGTATATTTCAGAACTGAACCATTTCAACCGAGCTGAAGCAT 50
D50092 rat        GACTCCTGAATATATTTCAAAACTGAACCATTTCAACCCAACTGAAGTAT 50
U78769 hamster    GACTCCTGAATATATTCCAAAACTGAACAATTTCAACTGAGCTGAAGTAC 50
Consensus         GACTCCTGAATATATTTCAAAACTGAACAATTTCAACCAAGCTGAAGCAT 50
D26150 cow        GACTTCTGAATATATTTGAAAACTGAACAGTTTCAACCAAGCCGAAGCAT 50
U67922 sheep      GACTTCTGAATATATTTGAAAACTGAACAGTTTCAACCAAGCTGAAGCAT 50
U29185 human      GACTCCTGAATATTTTTCAAAACTGAACAATTTCAGCCATGTCTGAGCTT 50
                  **** **** *** **  * ********  ***** *        **
                                51 Conserved Residues

U29186 mouse      TCTGCCTTCCTAGTGGTACCAGTCCAATTT-AGGAGAGCCA-AGCAGACT 98
D50092 rat        TCTGCCTTCTTAGCGGTACCAGTCCGGTTT-AGGAGAGCCA-AGCCGACT 98
U78769 hamster    TCTGTTTTTCTAGAGGTACCAGTTCAGTTT-AGGAGAGTCACAGCAGATC 99
Consensus         TCTGTCTTCCTAGAGGTACCAGTCCAGTTT-AGGAGAGCCACAGCAGATT 99
D26150 cow        -CTGTCTTCCCAGAGACACAAATCCAACTTGAGCTGAATCACAGCAGAT- 98
U67922 sheep      -CTGTCTTCCCAGAGACACAGATCCAACTTGAGCTGAATCACAGCAGAT- 98
U29185 human      TCCGTCTTCCTGGAGGCACAAATCTAGTTT-AGCTGAACCACAACAGATT 99
                   * *  **    * *  **   *     ** **  **  ** * * **


                 Consensus  human   cow   sheep   mouse   rat   hamster
Consensus        Ave:85%    81      81    81      90      88    89
human                       -       77    76      72      66    71
cow                                 -     97      73      58    75
sheep                                     -       73      58    76
mouse                                             -       89    82
rat                           % Identities                -     81
hamster                     Average:74.9%                         -
```

Sequenzlogo

- Zusammenfassung des Alignments vieler Sequenzen (http://weblogo.berkeley.edu/)

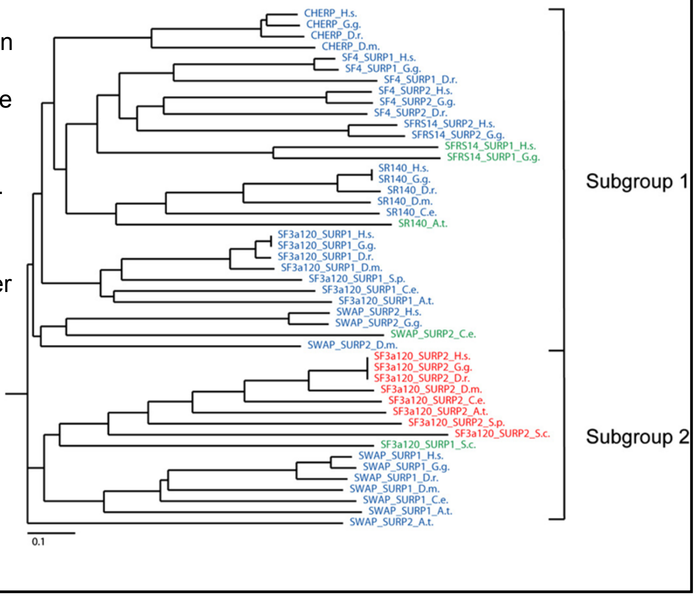


- Informationsgehalt (y-Achse): $R_i = \log_2 20 - H_i - e_n$
  ($e_n = \dfrac{20-1}{2n\ \ln 2}$; Korrektur für wenige Sequenzen *n*)
- Shannon-Entropie (Ungewissheit): $H_i = -\sum_{k=1}^{20} f_{ki} \log_2 f_{ki}$
  f_{ki}= relative Häufigkeit von AS k an Sequenzposition i
- Höhe der Buchstaben (AS-Codes): $f_{ki} R_i$

Phylogentische Bäume

Länge der horizontalen Linien entspricht der Anzahl Mutationen, die notwendig sind, um eine Sequenz in die andere zu überführen.

Ermöglicht Clustering in Gruppen verwandter Sequenzen.


CHERP_H.s.
CHERP_G.g.
CHERP_D.r.
CHERP_D.m.
SF4_SURP1_H.s.
SF4_SURP1_G.g.
SF4_SURP1_D.r.
SF4_SURP2_H.s.
SF4_SURP2_G.g.
SF4_SURP2_D.r.
SFRS14_SURP2_H.s.
SFRS14_SURP2_G.g.
SFRS14_SURP1_H.s.
SFRS14_SURP1_G.g.
SR140_H.s.
SR140_G.g.
SR140_D.r.
SR140_D.m.
SR140_C.e.
SR140_A.t.
SF3a120_SURP1_H.s.
SF3a120_SURP1_G.g.
SF3a120_SURP1_D.r.
SF3a120_SURP1_D.m.
SF3a120_SURP1_S.p.
SF3a120_SURP1_C.e.
SF3a120_SURP1_A.t.
SWAP_SURP2_H.s.
SWAP_SURP2_G.g.
SWAP_SURP2_C.e.
SWAP_SURP2_D.m.
SF3a120_SURP2_H.s.
SF3a120_SURP2_G.g.
SF3a120_SURP2_D.r.
SF3a120_SURP2_D.m.
SF3a120_SURP2_C.e.
SF3a120_SURP2_A.t.
SF3a120_SURP2_S.p.
SF3a120_SURP2_S.c.
SF3a120_SURP1_S.c.
SWAP_SURP1_H.s.
SWAP_SURP1_G.g.
SWAP_SURP1_D.r.
SWAP_SURP1_D.m.
SWAP_SURP1_C.e.
SWAP_SURP1_A.t.
SWAP_SURP2_A.t.
0.1
Subgroup 1
Subgroup 2