Aufbaukurs Bioinformatik

Folgende Techniken sollen erlernt werden:

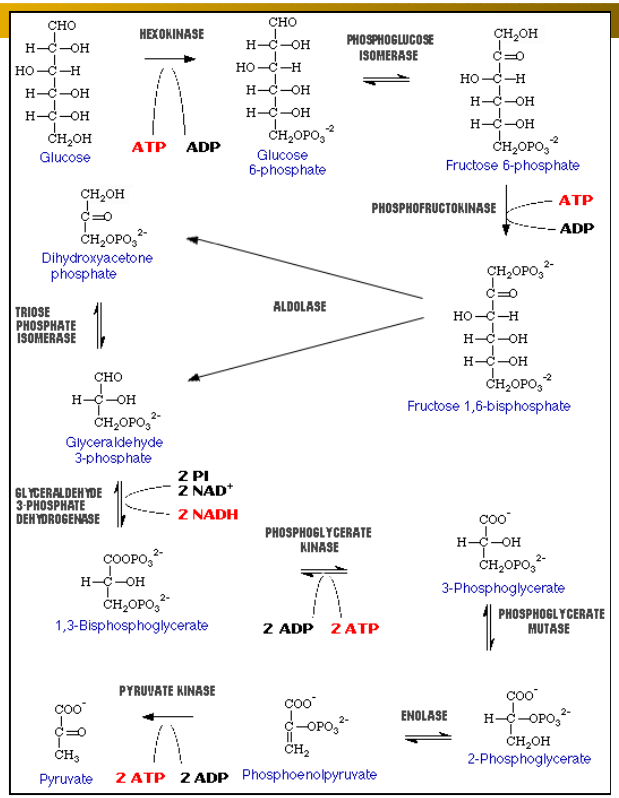
- Sequenzvergleich und Online-Resourcen hierfür
- Öffentliche Datenbanken zu Genen/Proteinen
- Strukturvorhersage
- Berechnung von Interaktionsnetzwerken

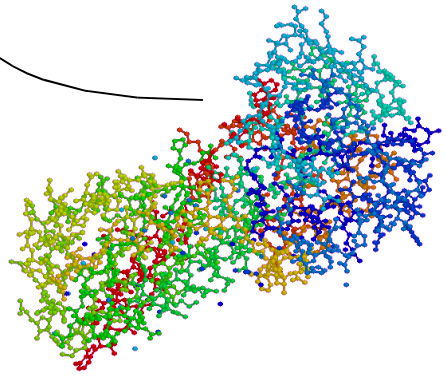
Ursula Kummer, Sven Sahle
mit Ulla Rost, Katja Wegner und Andreas Weidemann

Aufbaukurs Bioinformatik

1. Tag: Proteinsequenzen und Sequenzvergleich

Die Zelle - vom Gen zum Enzym





MFKPVDFSETSPVPPDIDLAPTQSPHHVAPSQDSSYDLLS……..
……. SMLKNKSFLLHGKDYPNNADNNDNEDIRAKTMNRSQSHV

gatccagctg taccattatg taatataata agacacggac gcac……..

Genetischer Code

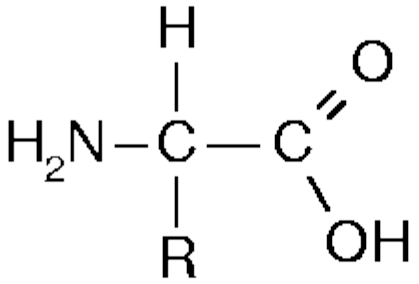
- Genetischer Code: Abbildung von Basen-Tripletts (Codone) auf Aminosäuren
  $\{XYZ \mid X,Y,Z \in \{A,C,G,U\} \rightarrow \{Ala, Cys, Asp, ..., Tyr\}$
- Eigenschaften des Genetischen Codes:
  - redundant (64 Codone werden auf 20 Aminosäuren und drei STOP-Codone abgebildet)
  - wird bei allen bekannten Lebewesen verwendet (mit leichten Abweichungen z.B. in Mitochondrien und in Prokaryonten, bisher 15 Code Tabellen bekannt)
  - fehlertolerant (gehört zu den 0.02 % Codierungen mit der höchsten Fehlertoleranz, d.h. Punktmutationen führen zu gleichen oder ähnlichen Aminosäuren)

Genetischer Code

P1	Position 2 G	A	C	U	P3
G	Gly	Glu	Ala	Val	G
	Gly	Glu	Ala	Val	A
	Gly	Asp	Ala	Val	C
	Gly	Asp	Ala	Val	U
A	Arg	Lys	Thr	Met	G
	Arg	Lys	Thr	Ile	A
	Ser	Asn	Thr	Ile	C
	Ser	Asn	Thr	Ile	U
C	Arg	Gln	Pro	Leu	G
	Arg	Gln	Pro	Leu	A
	Arg	His	Pro	Leu	C
	Arg	His	Pro	Leu	U
U	Trp	STOP	Ser	Leu	G
	STOP	STOP	Ser	Leu	A
	Cys	Tyr	Ser	Phe	C
	Cys	Tyr	Ser	Phe	U

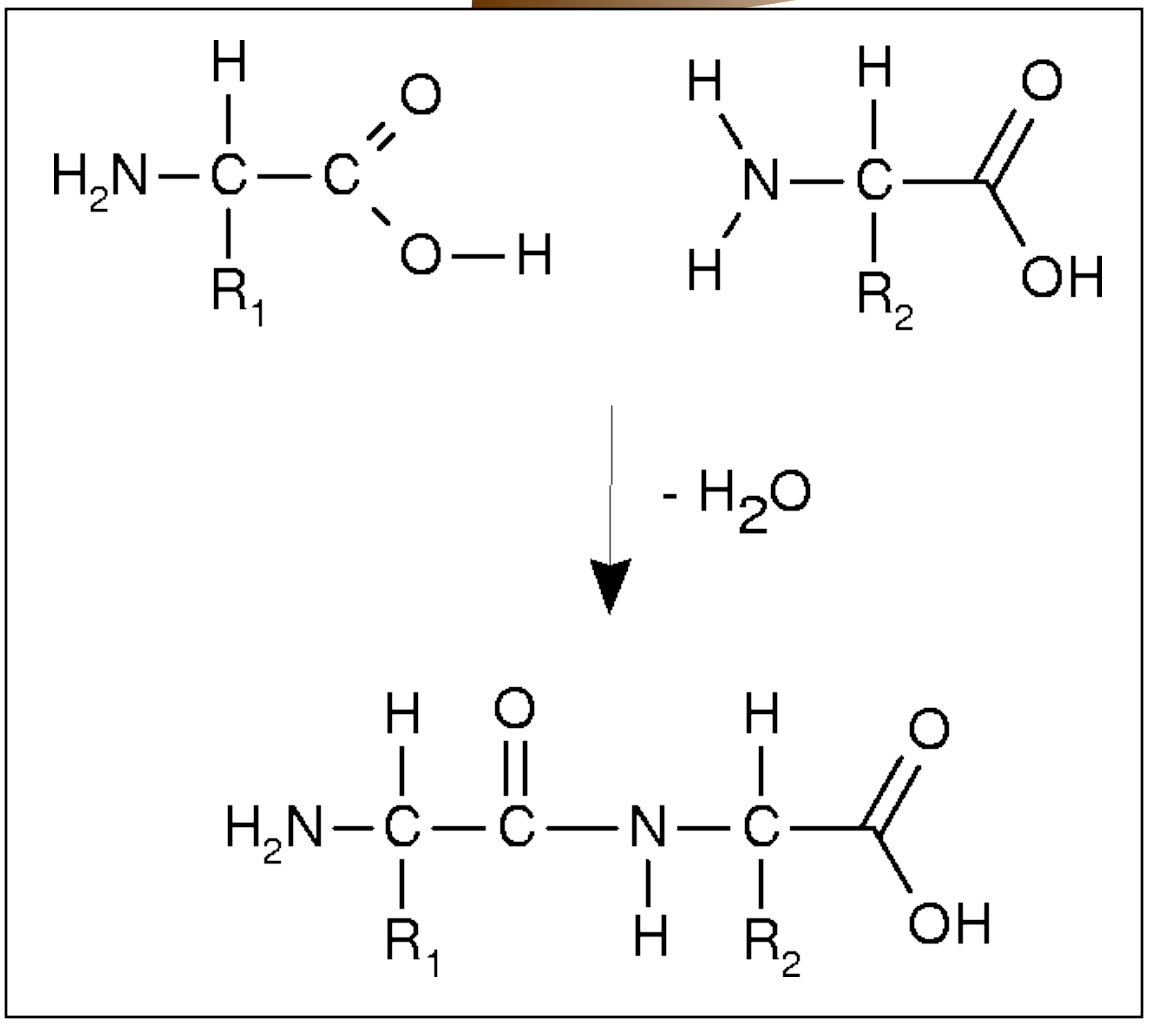
A	Ala	Alanin
C	Cys	Cystein
D	Asp	Aspartat
E	Glu	Glutatmat
F	Phe	Phenyalanin
G	Gly	Glycin
H	His	Histidin
I	Ile	Isoleucin
K	Lys	Lysin
L	Leu	Leucin
M	Met	Methionin
N	Asn	Asparagin
P	Pro	Prolin
Q	Gln	Glutamin
R	Arg	Arginin
S	Ser	Serin
T	Thr	Threonin
V	Val	Valin
W	Trp	Tryptophan
Y	Tyr	Tyrosin

Aminosäuren: Aufbau



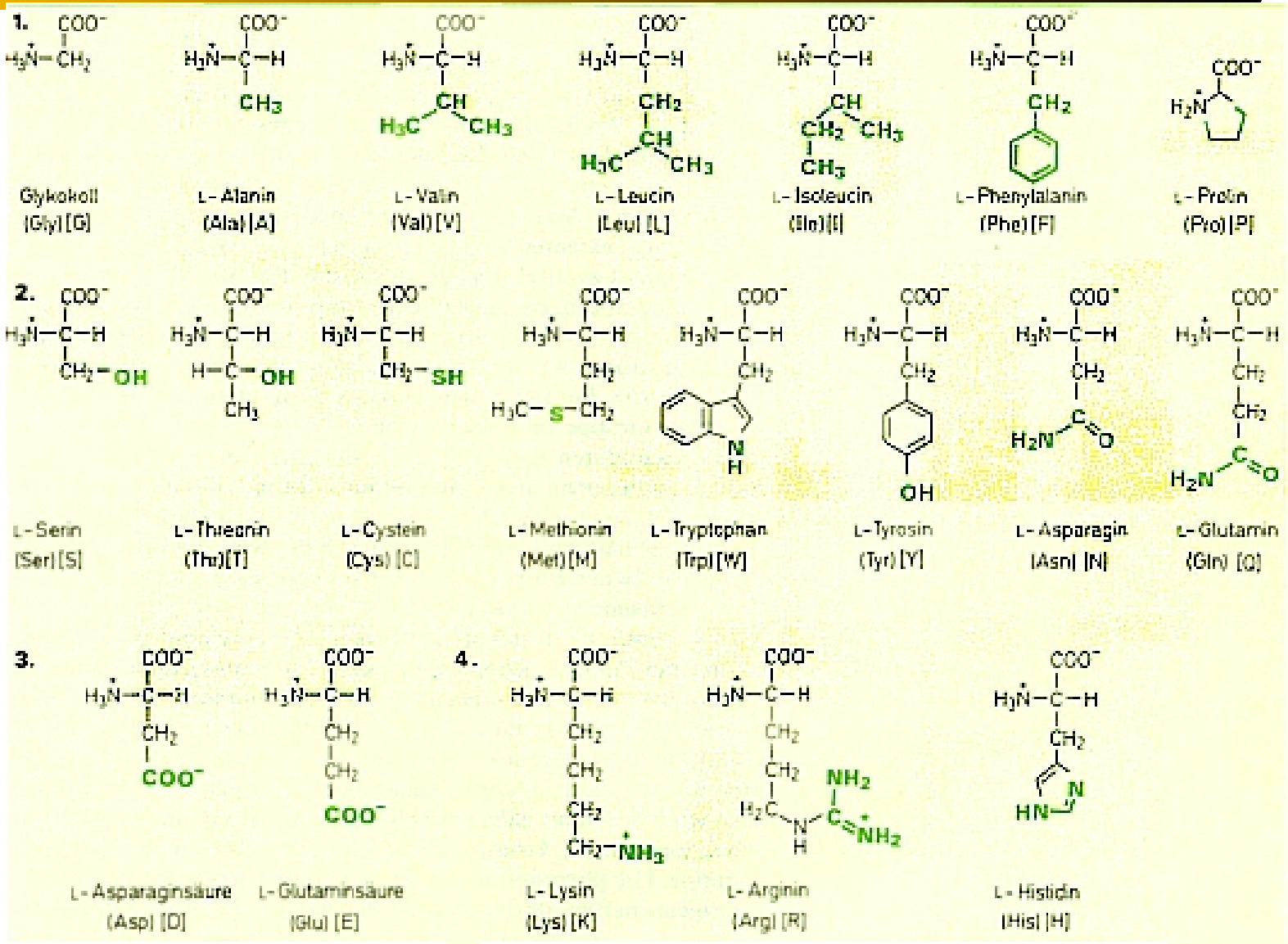
R: Rest

Struktur einer Aminosäure



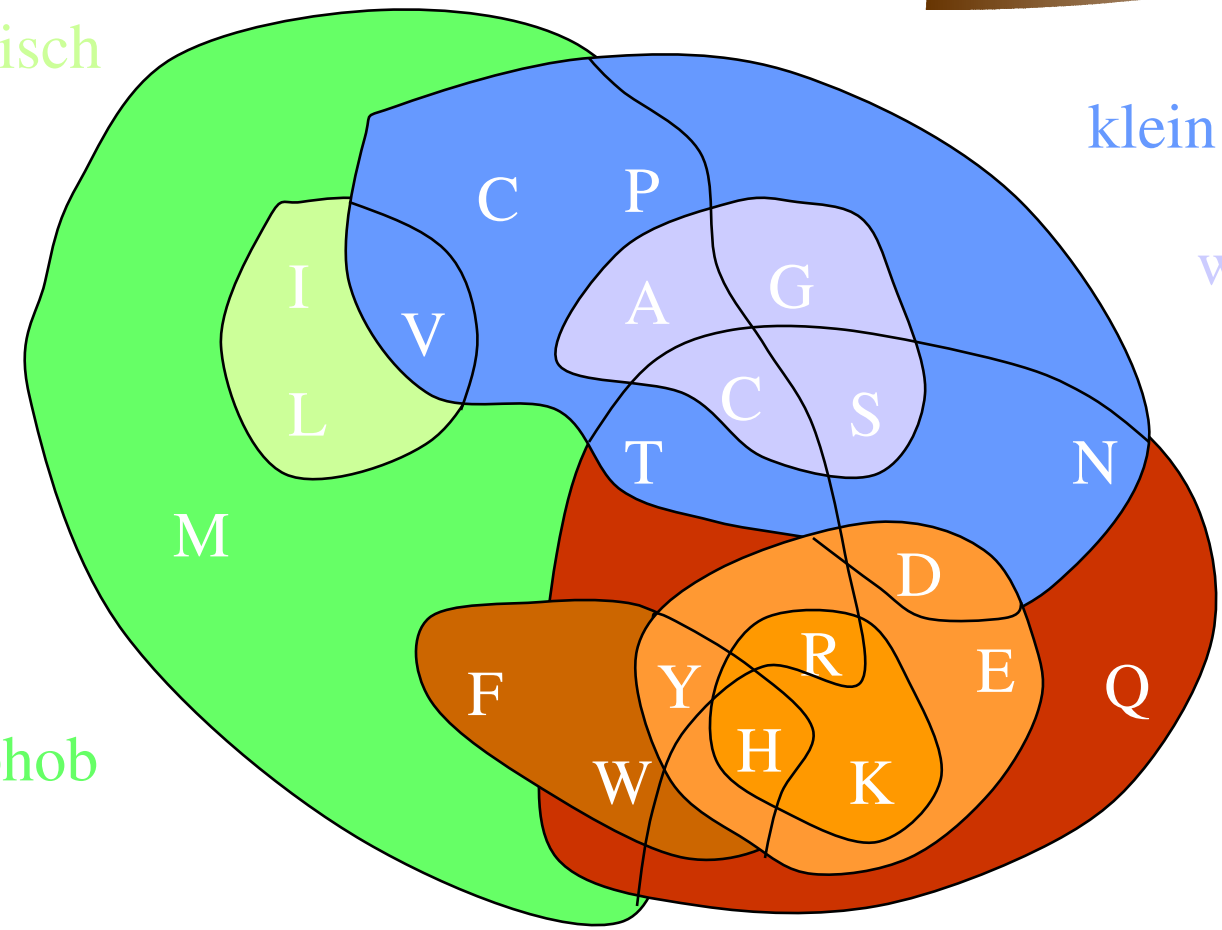
Entstehung einer Peptidbindung zwischen zwei Aminosäuren

Aminosäuren



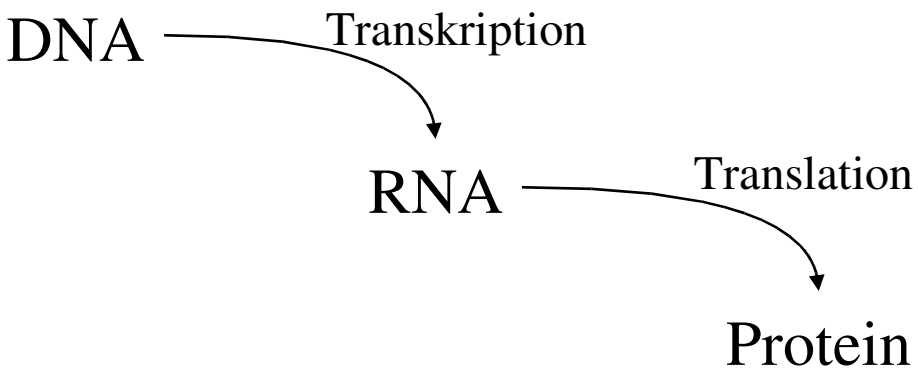
Quelle: lehrer2.rz.uni-karlsruhe.de/~mmb

Eigenschaften von Aminosäuren



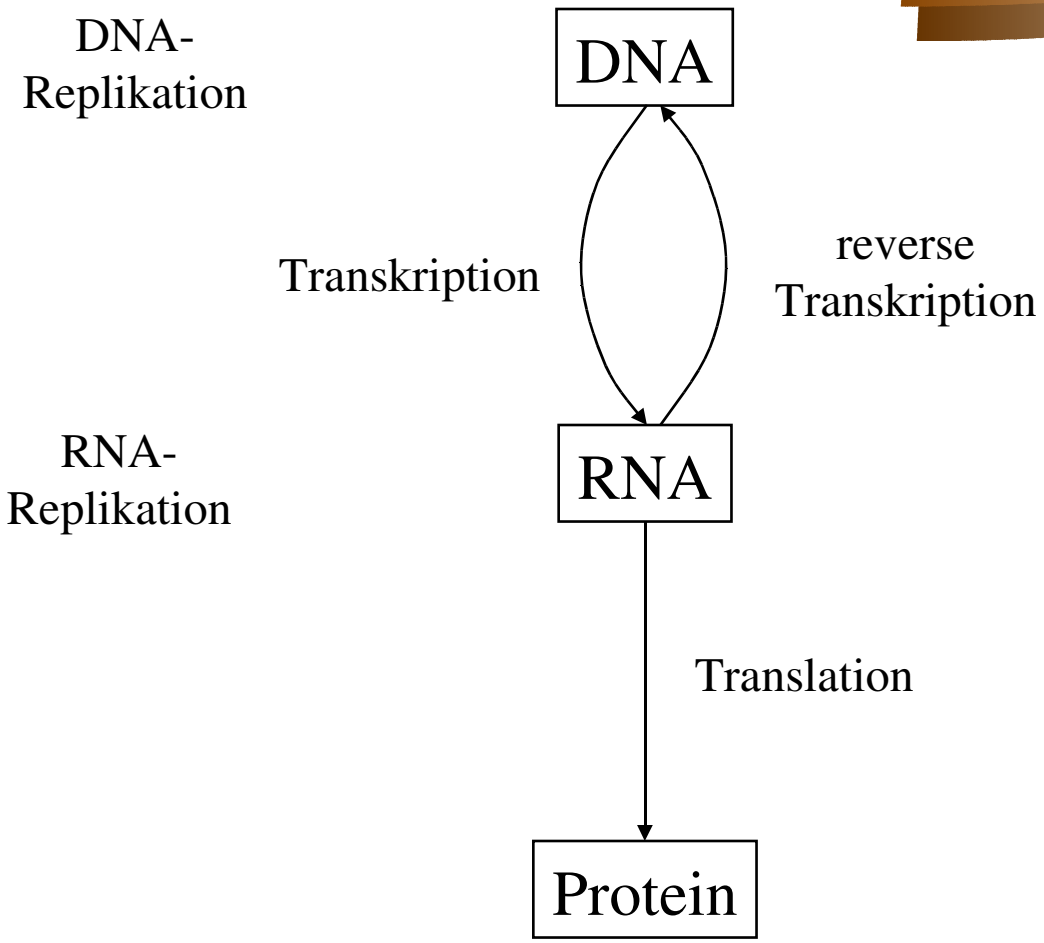
DNA → RNA → Protein

- Zentrales Dogma der Biochemie:
  - Der Fluß der Genetischen Information verläuft von der DNA zur RNA zum Protein



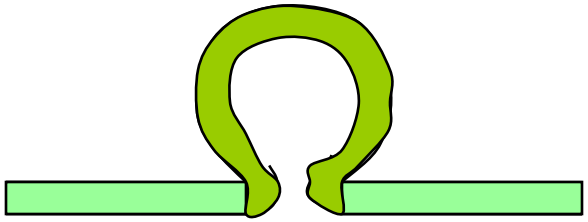
- 1964 kam die Hypothese auf, daß RNA Viren aus ihrer RNA wiederum DNA bilden können.
- 1970 wurde das verantwortliche Enzym gefunden. Folge: Die Lehrbücher mußten umgeschrieben werden

Erweitertes Dogma

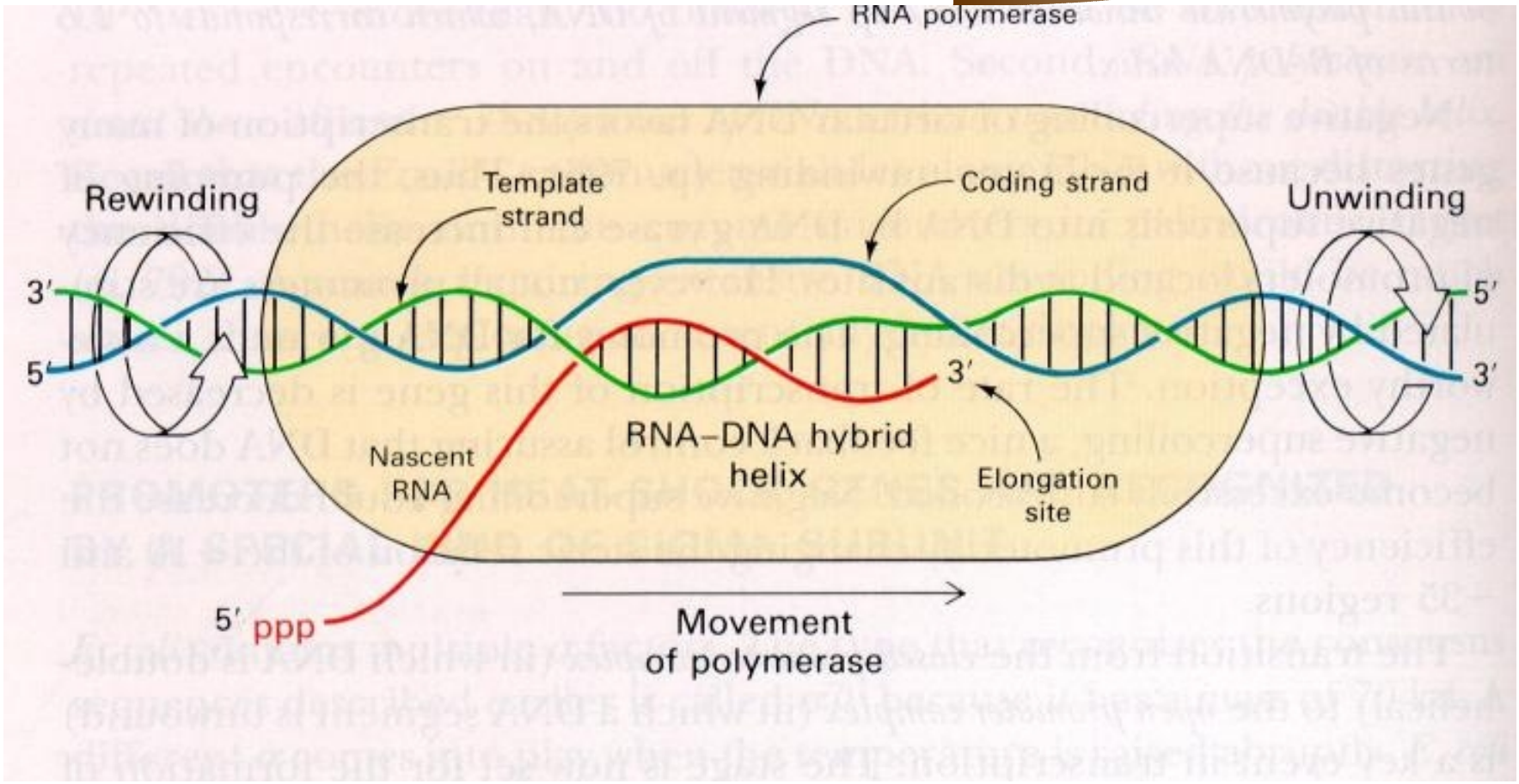


Transkription

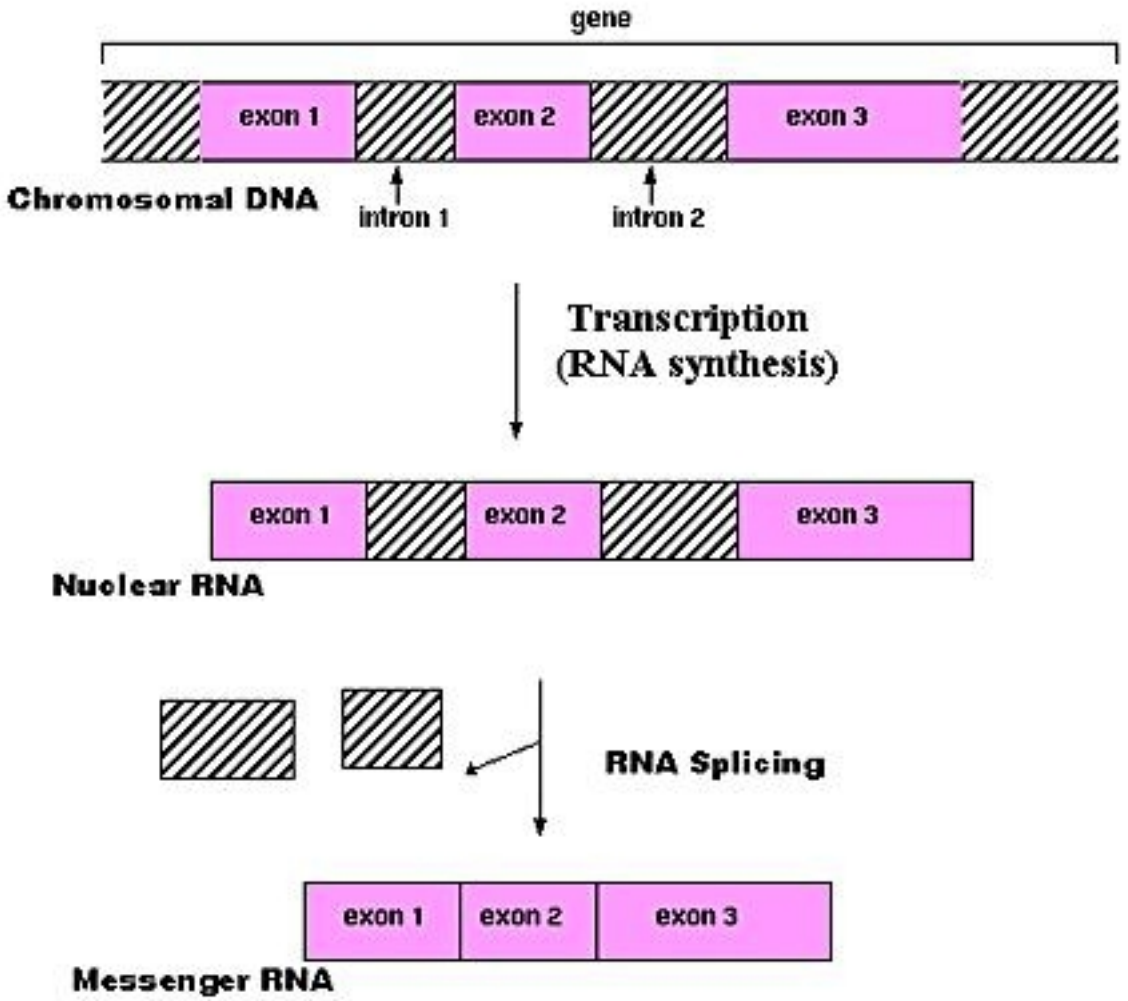
- Umschreibung von DNA in RNA
  - ablesen vom nicht codierenden Strang (3´→ 5´), aber erzeugt Strang (5´→ 3´),
  - ersetzen von Thymin durch Uracil
  - wird durch RNA-Polymerase katalysiert: jeweils auf kurzem Stück wird DNA in Einzelstränge aufgetrennt
  - Start bei Promoter-Sequenz
  - abspalten der RNA nach erreichen einer bestimmten Sequenz (z.B. ...AATAAA...)
  - danach Splicing der RNA:
    - rausschneiden der Introns
    - alternatives Splicen für unterschiedliche mRNA
  - ungefähr 30 Nukleotide pro Sekunde



Transkription



Intron-Extron

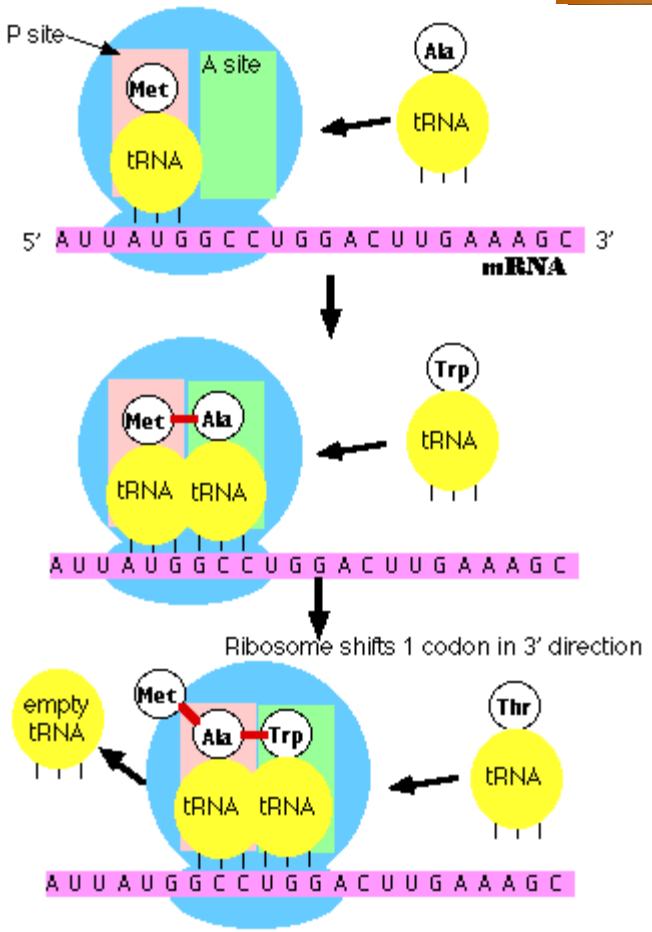


RNA synthesis and processing

Translation

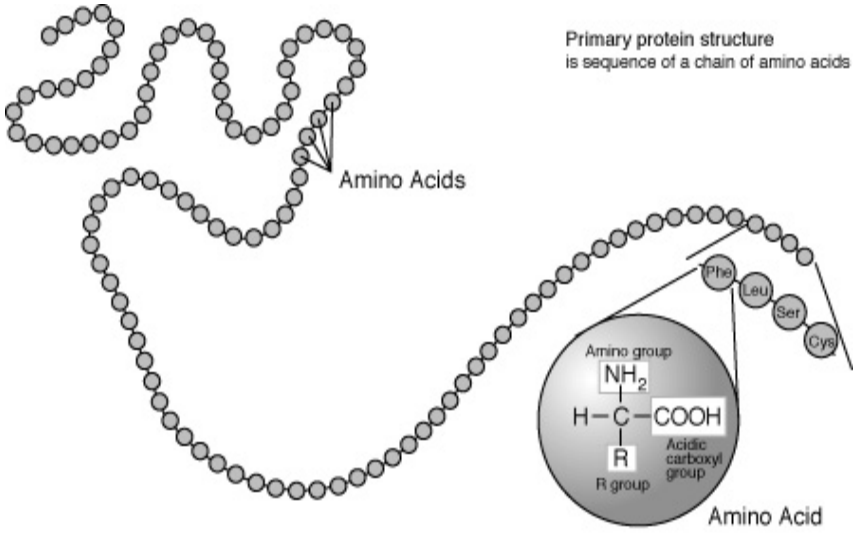
- Proteinsynthese im Ribosom
  - Teilnehmer: mRNA (bestimmen Reihenfolge der Aminosäuren), tRNA (transportieren Aminosäuren), Ribosome (Ort der Translation, Regulation der Bindung von mRNA und tRNA)
  - Initiation : Ribosom bindet an Start-Codon in mRNA 5´ AUG 3´
  - Aktivierung : tRNA bindet an Start-Codon in mRNA und ist assoziiert zu A-Untereinheit des Ribosoms
  - Elongation: tRNA wird “weitergereicht” an P-Untereinheit, nächste tRNA mit passendem Anti-Codon bindet an mRNA
  - Aminosäure der ersten tRNA bindet an Aminosäure der zweiten tRNA, Rest verläßt das Ribosom
  - Termination: Ribosom liest Stop-Codon (5´ TAA 3´, 5´ TAG 3´, 5´ TGA 3´)

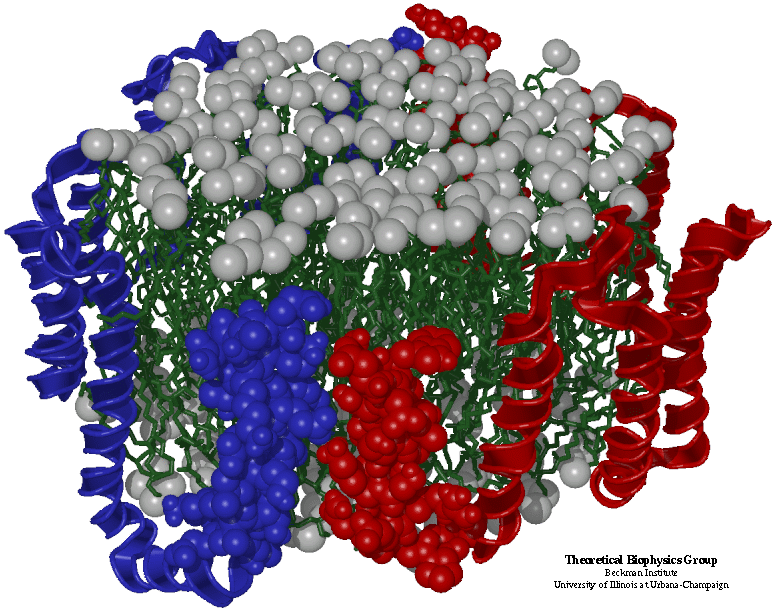
Translation



Proteine

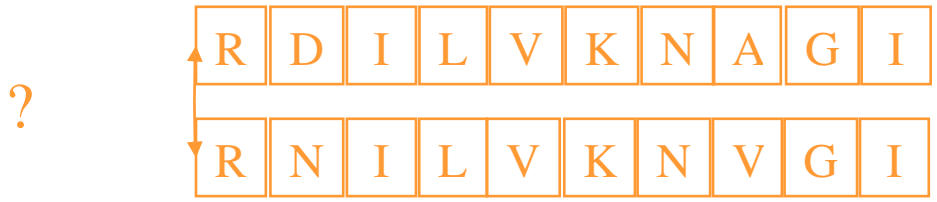
- Proteine sind Makromoleküle, die viele Abläufe in der Zelle bestimmen (Strukturbildung- und Erhaltung, Transport, Schutz und Abwehr, Steuerung und Regelung, Katalyse, Bewegung, Speicherung)
- Im menschlichen Körper werden etwa 100 000 verschiedene Proteine vermutet.





Paarweiser Sequenzvergleich

- Analyse der evolutionären Beziehung
- Ausgangspunkt für die Vorhersage der Funktion eines Proteines oder seiner Struktur (oder beides)



Ähnlichkeits-Dogma

- Wenn zwei Sequenzen sehr ähnlich sind, haben sie auch eine
  - eine ähnliche Funktion,
  - eine ähnliche Struktur,

  und sie haben einen gemeinsamen Vorfahren

  Vorsicht: das stimmt nicht immer !!!
- Das impliziert, daß
  - die Sequenz eine Syntax bildet, die eine Funktion codiert
  - es gibt auch Redundanz, da einige Elemente ausgetauscht werden können, ohne daß sich die Funktion ändert (robuste Semantik)

Vergleich von Proteinsequenzen

- Paarweiser Vergleich von je einem Buchstaben aus zwei Sequenzen, d.h. keine Betrachtung statistischer Abhängigkeiten innerhalb einer Sequenz
- Ähnlichkeit von zwei Sequenzen ergibt sich als Summe aus den Einzelähnlichkeiten (Markov-Modell)
- Aufstellung von sogenannten Scoring-Matrizen: Ähnlichkeitswert bezieht sich immer nur auf das dahinterliegende Modell
- Verfahren hauptsächlich für den Vergleich von Aminosäuresequenzen (Proteinen)

Vergleich von Aminosäuren

- Einfachste Vergleichsmöglichkeit: Identitäts-Matrix
  - gleiche Buchstaben = 1,
  - ungleiche Buchstaben = 0
- Ähnlichkeitsmaße, die über = / ≠ Vergleiche hinausgehen, nutzen
  - chemische oder strukturelle Eigenschaften: polar/unpolar, Form, Größe, Ladung
  - genetische Eigenschaften: minimale Anzahl ausgetauschter Basen in der dazugehörigen DNA
  - evolutionäre Distanz: beobachtete Austausch-Häufigkeiten von Aminosäuren (in bekannten Proteinfamilien)

Identitätsmatrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	1																			
R	0	1																		
N	0	0	1																	
D	0	0	0	1																
C	0	0	0	0	1															
Q	0	0	0	0	0	1														
E	0	0	0	0	0	0	1													
G	0	0	0	0	0	0	0	1												
H	0	0	0	0	0	0	0	0	1											
I	0	0	0	0	0	0	0	0	0	1										
L	0	0	0	0	0	0	0	0	0	0	1									
K	0	0	0	0	0	0	0	0	0	0	0	1								
M	0	0	0	0	0	0	0	0	0	0	0	0	1							
F	0	0	0	0	0	0	0	0	0	0	0	0	0	1						
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1					
S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1				
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1			
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1		
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z	
3.0	2.0	1.0	2.0	2.0	1.0	2.0	1.0	1.0	1.0	1.0	1.0	1.0	2.0	1.0	1.0	2.0	2.0	2.0	1.0	1.0	2.0	A
	3.0	1.0	3.0	2.0	1.0	2.0	2.0	2.0	2.0	1.0	1.0	3.0	1.0	2.0	1.0	2.0	2.0	2.0	0.0	2.0	2.0	B
		3.0	1.0	0.0	2.0	2.0	1.0	1.0	0.0	1.0	0.0	1.0	1.0	0.0	2.0	2.0	1.0	1.0	2.0	2.0	0.0	C
			3.0	2.0	1.0	2.0	2.0	1.0	1.0	1.0	0.0	2.0	1.0	1.0	1.0	1.0	1.0	2.0	0.0	2.0	2.0	D
				3.0	0.0	2.0	1.0	1.0	2.0	1.0	1.0	1.0	1.0	2.0	1.0	1.0	1.0	2.0	1.0	1.0	3.0	E
					3.0	1.0	1.0	2.0	0.0	2.0	1.0	1.0	1.0	0.0	1.0	2.0	1.0	2.0	1.0	2.0	0.0	F
						3.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	2.0	2.0	1.0	2.0	2.0	1.0	2.0		G
							3.0	1.0	1.0	2.0	0.0	2.0	2.0	2.0	2.0	1.0	1.0	1.0	0.0	2.0	2.0	H
								3.0	2.0	2.0	2.0	2.0	1.0	1.0	2.0	2.0	2.0	2.0	0.0	1.0	1.0	I
									3.0	1.0	2.0	2.0	1.0	2.0	2.0	1.0	2.0	1.0	1.0	1.0	2.0	K
										3.0	2.0	1.0	2.0	2.0	2.0	2.0	1.0	2.0	2.0	1.0	2.0	L
											3.0	1.0	1.0	1.0	2.0	1.0	2.0	2.0	1.0	0.0	1.0	M
												3.0	1.0	1.0	1.0	2.0	2.0	1.0	0.0	2.0	2.0	N
													3.0	2.0	2.0	2.0	2.0	1.0	1.0	1.0	2.0	P
														3.0	2.0	1.0	1.0	1.0	1.0	1.0	3.0	Q
															3.0	2.0	2.0	1.0	2.0	1.0	2.0	R
																3.0	2.0	1.0	2.0	2.0	1.0	S
																	3.0	1.0	1.0	1.0	1.0	T
																		3.0	1.0	1.0	2.0	V
																			3.0	1.0	1.0	W
																				3.0	1.0	Y
																					3.0	Z

Genetische Matrix

B = D∨ N

Z = E ∨ Q

PAM-Matrix

- basiert auf evolutionärem Modell
  - ähnliche Proteine haben einen gemeinsamen Vorfahren, aus dem beide Sequenzen durch genetische Veränderungen wie z.B. Punktmutationen hervorgegangen sind (Edit-Distanz)
- empirisch aus Vorkommen von Aminosäuren in ähnlichen (mindestens 85% identischen), homologen Proteinen abgeschätzt
- PAM : Accepted Point Mutation
- PAM 1 - Matrix
  - 1 evolutionärer Schritt
  - 1 Mutation pro 100 Residuen erlaubt (1% Unterschied)
  - wie hoch ist Wahrscheinlichkeit, daß sich ein Residuum ändert?

Berechnung von PAM

1. Schritt: Evolutionäres Modell aufstellen
2. Schritt: Häufigkeiten der Aminosäuren bestimmen
3. Schritt: Mutationshäufigkeit jeder Aminosäure bestimmen
4. Schritt: Matrix mit Austauschwahrscheinlichkeiten bestimmen
5. Schritt: Evolutionäre Skalierung
6. Schritt: Relative Wahrscheinlichkeit
7. Schritt: Log-odds

Schritt 6

M gibt die absolute Wahrscheinlichkeit eines Austauschs an.

Selbst bei zufälligen Sequenzen erhält man aber eine Übereinstimmung von etwa 5%.

Daher müssen die Werte für einen Austausch einer Aminosäure noch in Relation zu dem Wert für einen zufälligen Austausch gesetzt werden.

$$p_i^{random} = f_i$$

Wahrscheinlichkeit für einen zufälligen Austausch

$$R_{i,j} = \frac{M_{i,j}}{p_i^{random}}$$

Matrix mit relativen Wahrscheinlichkeiten

Alternative Scoring-Matrizen

- Zahlreiche andere Matrizen
  - anders normalisiert (GCG Version von MDM78 PAM250)
  - neu berechnet aus neueren und umfangreicheren Daten (PET)
  - mit Hilfe von großen Datenmengen (1.7Mio) direkt für größere evolutionäre Distanzen berechnet (GCB, G. Gonnet, M.A. Cohen, & S. Benner (1992))
  - abgeleitet aus Blöcken hoch-konservierter Bereiche in den Proteinen, die mit einem Schwellwert geclustert werden (BLOSUM-t)
  - aus Vergleich der Tertiärstruktur bei Proteinen mit bekannter 3D-Struktur andere Austausch-Wahrscheinlichkeiten abgeleitet

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

BLOSUM 62

X : any

Beispiel

R	D	I	L	V	K	N	A	G	I
R	N	I	L	V	K	N	V	G	I

Identitäts-Matrix : 1 + 0 + 1 + 1 + 1 + 1 + 1 + 0 + 1 + 1 = 8

Genetische Matrix: 3 + 2 + 3 + 3 + 3 + 3 + 3 + 2 + 3 + 3 = 28

PAM250: 6 + 2 + 5 + 6 + 4 + 5 + 2 + 0 + 5 + 5 = 40

BLOSUM62: 5 + 1 + 4 + 4 + 4 + 5 + 6 + 0 + 6 + 4 = 39

Auswahl der Matrix

- Hängt stark von den zu vergleichenden Sequenzen ab
  - je nach (bekanntem)Verwandtschaftsgrad 120 PAM - 250 PAM
- BLOSUM oft recht gut bei bestimmten Suchverfahren (z.B. BLAST), aber nicht durchgängig für alle Protein-Familien

Datenbanksuche

- **Ziel**: Aus einer Sequenz-Datenbank alle Einträge bestimmen, die ähnlich zu einer vom Benutzer vorgegebenen Query-Sequenz sind
- **Problem**: Aufwand für ein optimales paarweises Alignment der Query-Sequenz gegen alle Datenbankeinträge viel zu hoch
- **Ansatz**: Heuristische Verfahren
  - BLAST
  - FASTA
- Versionen für Nukleotide und DNA

BLAST

- Basic Local Alignment Search Tool
- generiert eine Liste von Segment-Paaren (Teilsequenzen gleicher Länge ohne Gaps) zwischen der Query-Sequenz und Einträgen der Datenbank, die einen Score oberhalb einer vorgegebenen Schwelle besitzen

BLAST

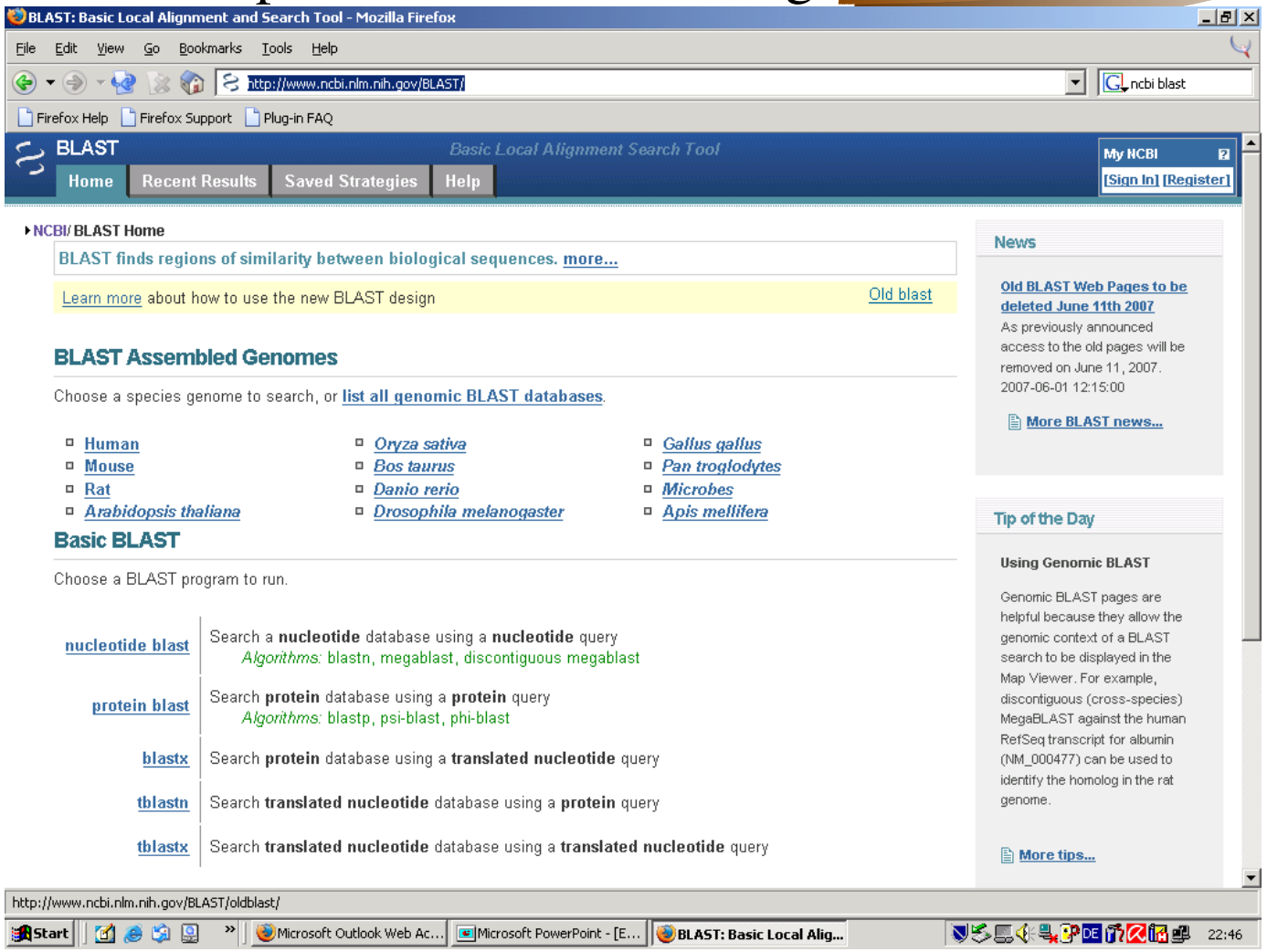
- Ausgangspunkt: Query-Sequenz, Wortlänge w, Schwellwerte *S,T*
- Drei-schrittiger Algorithmus:
  - für eine vorgegebene Wort-Länge *w* und eine Score-Matrix werden alle Wörter der Länge *w* bestimmt, die bei einem Vergleich mit der Query-Sequenz einen Score > *T* ergeben würden
  - Die Datenbank wird auf diese sog. w-Mere hin durchsucht
  - Jeder Treffer (Sequenz aus DB) wird in beide Richtungen erweitert (ohne Gaps) und es wird geprüft, ob sich ein Score > *S* ergibt
  - Ausgabe aller Segmente mit einem Score > *S*

BLAST

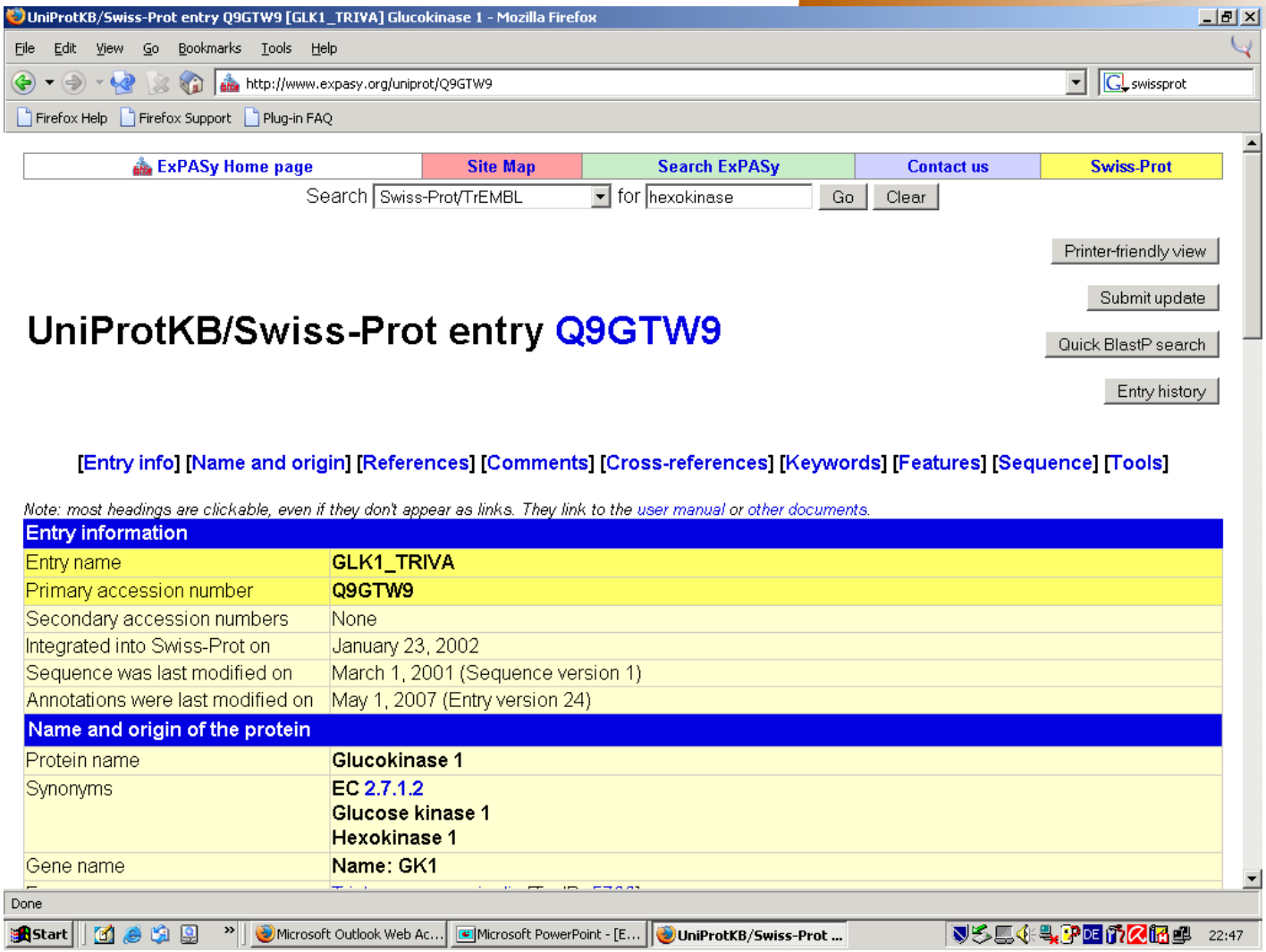
- Varianten:
  - BLASTn : für Nukleotidsequenzen (DNA)
  - BLASTp: für Aminosäuresequenzen (Proteine)
  - BLASTx: macht eine 6-Frame Translation

BLAST: Beispiel

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

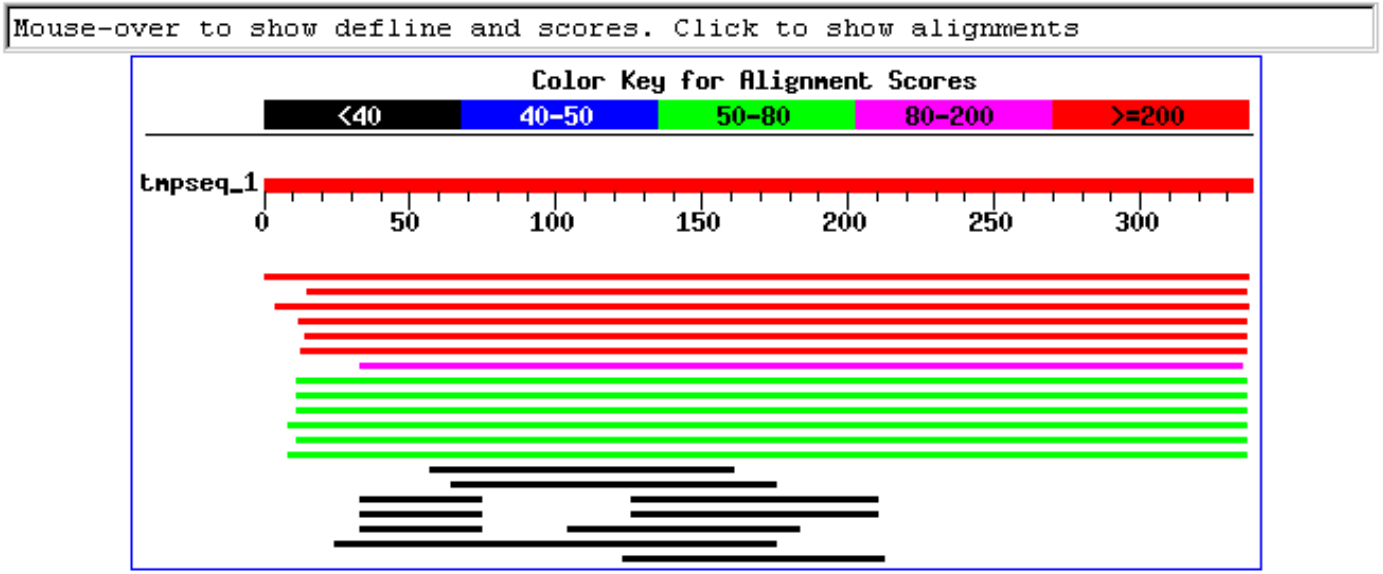


BLAST: Beispiel



BLAST: Beispiel

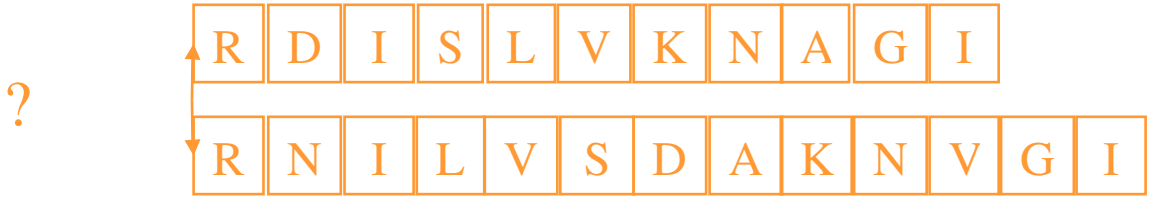
Distribution of 23 Blast Hits on the Query Sequence



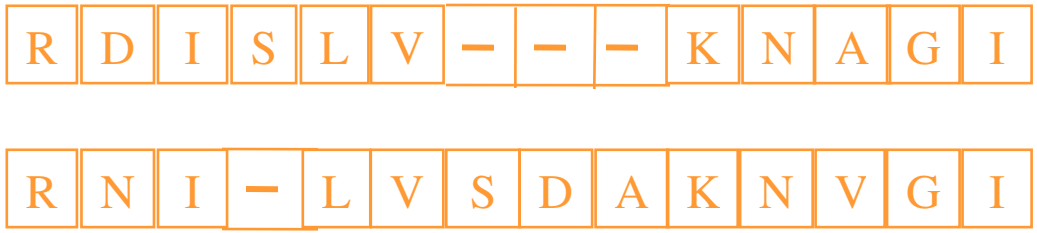
Sequences producing significant alignments:	Score (bits)	E Value
ref|NP_011083.1| Yer156cp >gi|731528|sp|P40093|YEY6_YEAST H...	660	0.0
emb|CAB71842.1| (AL138666) conserved hypothetical protein [...	309	2e-83
gb|AAF64518.1|AF252871_1 (AF252871) GAMM1 protein [Mus musc...	262	3e-69
pir||T19538 hypothetical protein KO8H10.8 - Caenorhabditis ...	257	1e-67
dbj|BAB08430.1| (AB017067) GAMM1 protein-like [Arabidopsis ...	255	6e-67

Gaps

- Neben Substitutionen können auch Einfügungen und Löschungen vorkommen



Alignment



Gap Penalty

- Gaps in einem Alignment (Paarung einer Aminosäure mit einer Lücke) werden mit einem schlechten Wert bestraft, z.B. negativer Wert für jedes
  - einzelne Gap, d.h. Funktion linear in der Länge *k* der eingefügten bzw. gelöschten Elemente

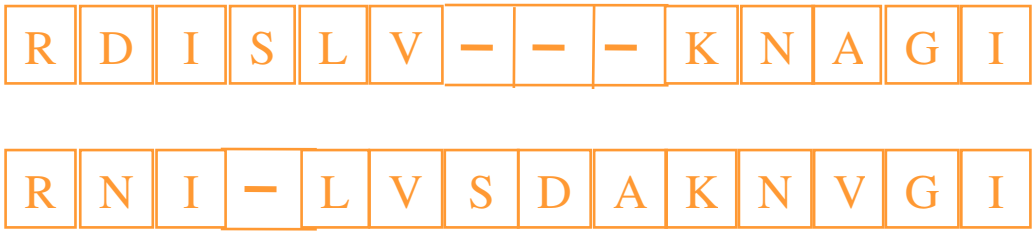
$$g(k)=q\cdot k$$

  - oder zusammengesetzter Wert für Einfügungen/Auslassungen beliebiger Länge
    a : Gap Eröffnungsstrafe b: Gap Ausweitungsstrafe

$$g(k)=a+b\cdot k$$

affin-lineare Gap Penalty

Beispiel



Identitäts-M. : 1 + 0 + 1 - g(1) + 1 + 1 - g(3) + 1 + 1 + 0 + 1 + 1

PAM250 (*10) : 6 + 2 + 5 - g(1) + 6 + 4 - g(3) + 5 + 2 + 0 + 5 + 5

Bioinformatik & Internet

- In den Anfangsjahren: Publikation aller Daten (Sequenzen, Gene, Proteine) in Journalen
- Mittlerweile: Speicherung der Daten in meist öffentlich zugänglichen Datenbanken
- Datenbanken oft sehr einfach strukturiert
- Ziel heute: Integration der Daten durch entsprechende Werkzeuge (Java, Perl,...) über das Netz

Datenbanken im Netz

- Gensequenzen (DNA + RNA): Kollaboration zwischen drei, weltweit verteilten Instituten mit täglichen Austausch der Daten (seit 1980)
  - GenBank at NCBI (National Center for Biotechnology Information) www.ncbi.nlm.nih.gov
  - European Molecular Biology Laboratory (EMBL) www.embl.de
  - DNA databank of Japan (DDBJ) www.ddbj.nig.ac.jp

NCBI

GenBank Overview - Microsoft Internet Explorer

Datei Bearbeiten Ansicht Favoriten Extras ?

Zurück Suchen Favoriten Medien

Adresse http://www.ncbi.nlm.nih.gov/Genbank/index.html Wechseln zu Links

NCBI **GenBank Overview**

PubMed Entrez BLAST OMIM Books Taxonomy Structure

Search Entrez for Go

NCBI

SITE MAP

Submit to GenBank

Updates

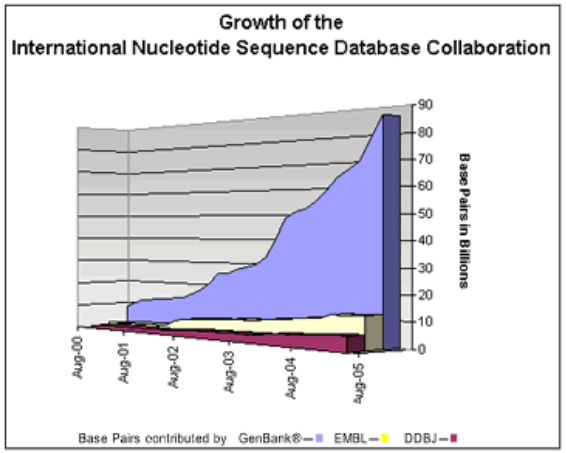
Search GenBank

Entrez Nucleotide

BLAST

International sequence databases exceed 100 gigabases

In August 2005, the INSDC announced the DNA sequence database exceeded 100 gigabases. GenBank is proud of its contributions toward this milestone. We thank all the scientists who have worked through the submission process at GenBank and made their sequence data available to the world. See the related press release.

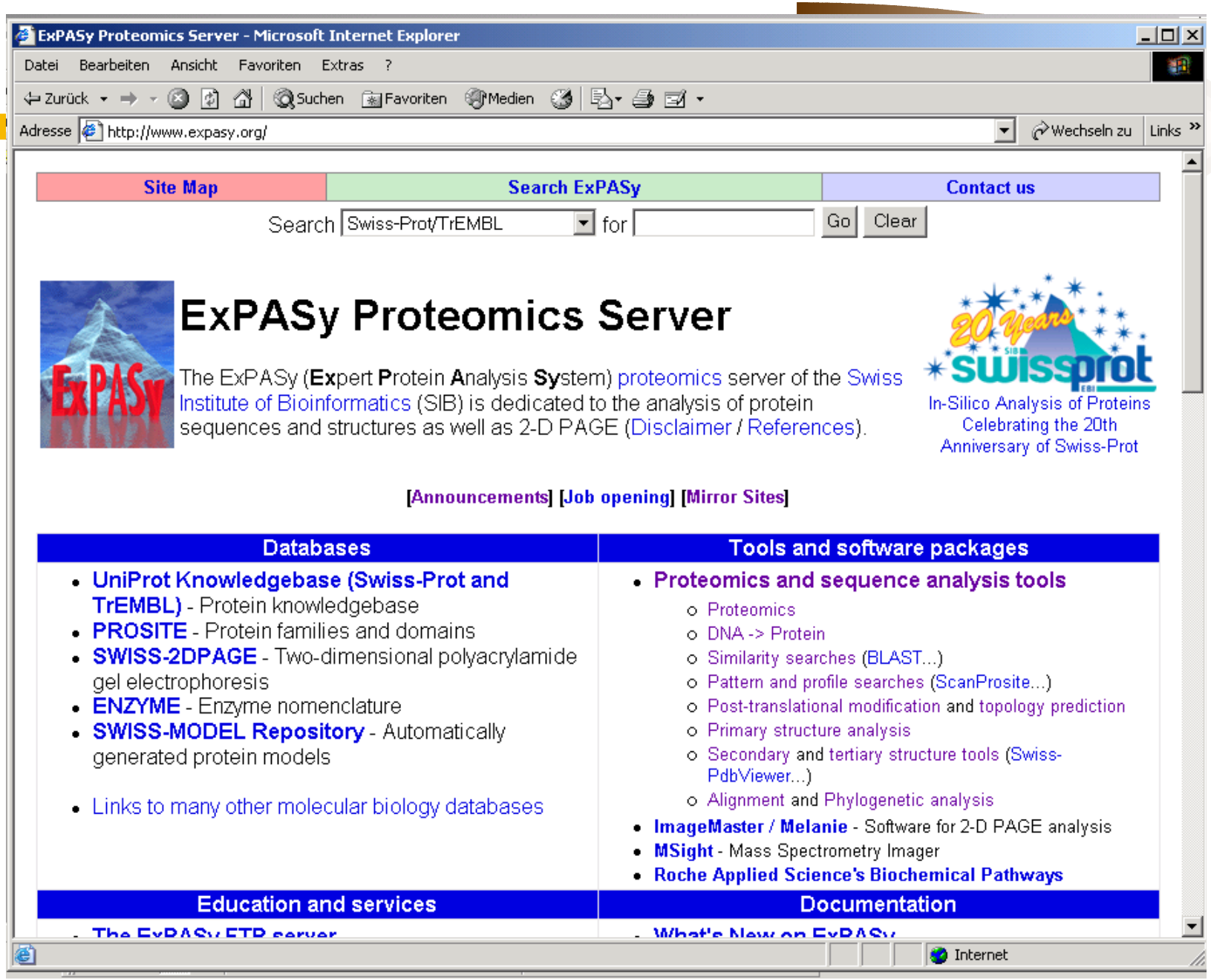


Internet

Datenbanken im Netz

- Protein Datenbanken
  - SWISSPROT http://www.expasy.org
  - PIR http://pir.georgetown.edu/

- Suche nach Proteinen
- Muster- und Sequenzvergleich
- Querreferenzen zu anderen Datenbanken
- Literaturreferenzen
- annotierte (und geprüfte Information)

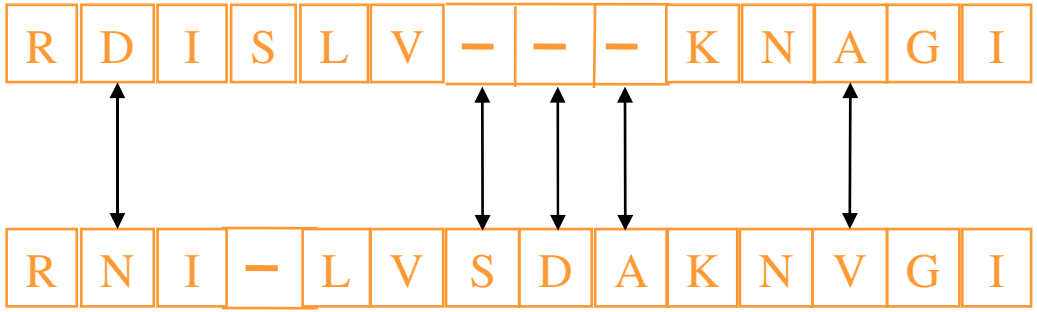

ExPASy Proteomics Server - Microsoft Internet Explorer
Datei Bearbeiten Ansicht Favoriten Extras ?
Zurück Suchen Favoriten Medien
Adresse http://www.expasy.org/
Wechseln zu Links
Site Map
Search ExPASy
Contact us
Search Swiss-Prot/TrEMBL for Go Clear
ExPASy Proteomics Server
The ExPASy (Expert Protein Analysis System) proteomics server of the Swiss Institute of Bioinformatics (SIB) is dedicated to the analysis of protein sequences and structures as well as 2-D PAGE (Disclaimer / References).
20 Years swissprot
In-Silico Analysis of Proteins Celebrating the 20th Anniversary of Swiss-Prot
[Announcements] [Job opening] [Mirror Sites]
Databases
UniProt Knowledgebase (Swiss-Prot and TrEMBL) - Protein knowledgebase
PROSITE - Protein families and domains
SWISS-2DPAGE - Two-dimensional polyacrylamide gel electrophoresis
ENZYME - Enzyme nomenclature
SWISS-MODEL Repository - Automatically generated protein models
Links to many other molecular biology databases
Tools and software packages
Proteomics and sequence analysis tools
Proteomics
DNA -> Protein
Similarity searches (BLAST...)
Pattern and profile searches (ScanProsite...)
Post-translational modification and topology prediction
Primary structure analysis
Secondary and tertiary structure tools (Swiss-PdbViewer...)
Alignment and Phylogenetic analysis
ImageMaster / Melanie - Software for 2-D PAGE analysis
MSight - Mass Spectrometry Imager
Roche Applied Science's Biochemical Pathways
Education and services
Documentation
Internet

SWISS-PROT Datenbankeintrag

```
ID   C5AR_HUMAN     STANDARD;      PRT;   350 AA.
AC   P21730;
DT   01-MAY-1991 (Rel. 18, Created)
DT   01-MAY-1991 (Rel. 18, Last sequence update)
DT   15-JUL-1998 (Rel. 36, Last annotation update)
DE   C5A ANAPHYLATOXIN CHEMOTACTIC RECEPTOR (C5A-R) (CD88 ANTIGEN).
GN   C5R1 OR C5AR.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Mammalia;
OC   Eutheria; Primates; Catarrhini; Hominidae; Homo.
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE; 91156029.
RA   GERARD N.P., GERARD C.;
RT   "The chemotactic receptor for human C5a anaphylatoxin.";
RL   Nature 349:614-617(1991).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE; 91175748.
RA   BOULAY F., MERY L., TARDIF M., BROUCHON L., VIGNAIS P.;
RT   "Expression cloning of a receptor for C5a anaphylatoxin on
RT   differentiated HL-60 cells.";
RL   Biochemistry 30:2993-2999(1991).
CC   -!- FUNCTION: RECEPTOR FOR THE CHEMOTACTIC AND INFLAMMATORY
PEPTIDE
CC       ANAPHYLATOXIN C5A. THIS RECEPTOR STIMULATES CHEMOTAXIS,
GRANULE
CC       ENZYME RELEASE AND SUPEROXIDE ANION PRODUCTION.
CC   -!- SUBCELLULAR LOCATION: INTEGRAL MEMBRANE PROTEIN.
CC   -!- SIMILARITY: BELONGS TO FAMILY 1 OF G-PROTEIN COUPLED
RECEPTORS.
CC   -!- DATABASE: NAME=PROW; NOTE=CD guide CD88 entry;
CC       WWW="http://www.ncbi.nlm.nih.gov/prow/cd/cd88.htm".
CC
---------------------------------------------------------------------
```

Optimales Paarweises Alignment

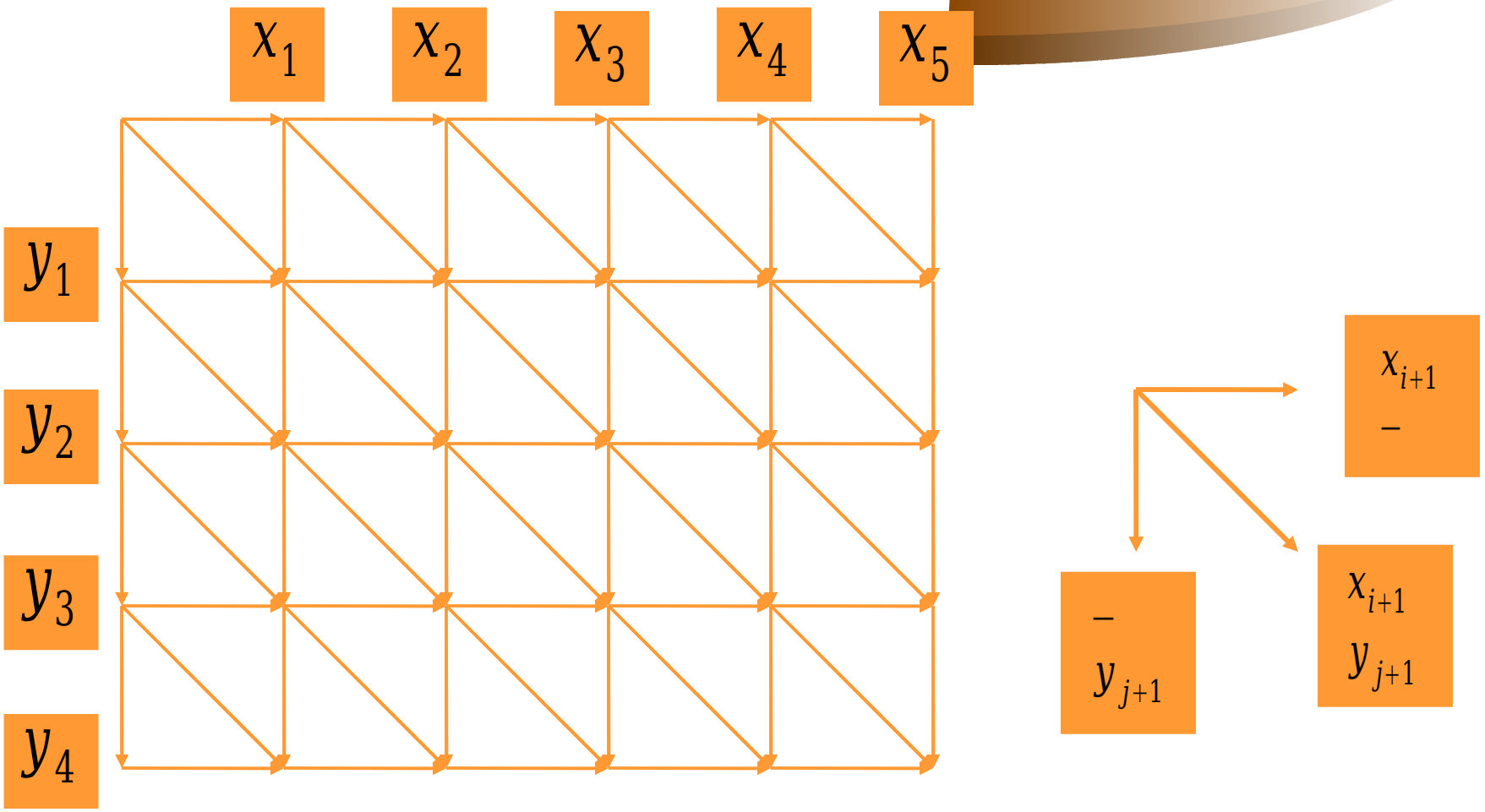
- Suche nach einem Alignment von zwei Sequenzen, so daß der Ähnlichkeitswert maximal ist (oder die Distanz minimal)



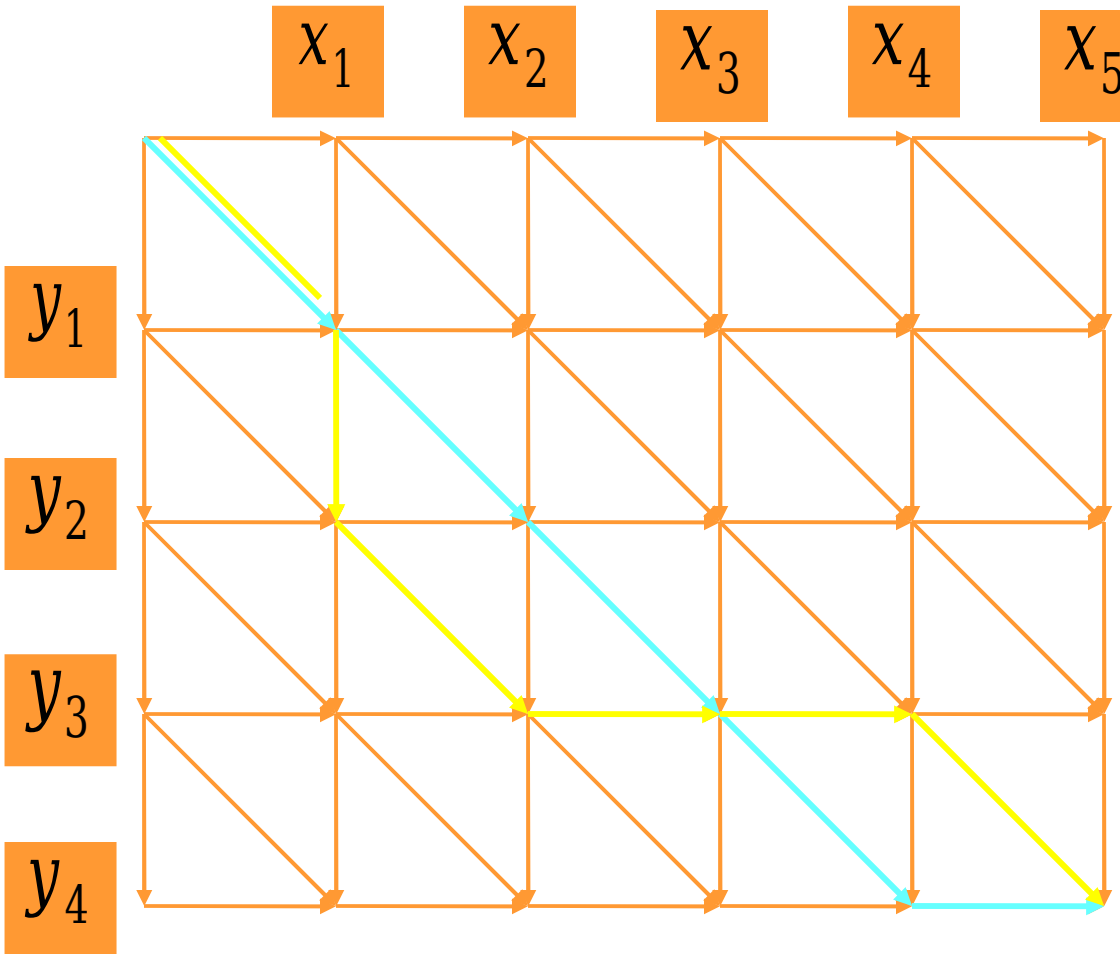
$$\binom{2n}{n} = \frac{(2n)!}{(n!)^2} \approx \frac{2^{2n}}{\sqrt{2\pi n}}$$

Möglichkeiten für zwei Sequenzen der Länge n

Alignment



Alignment



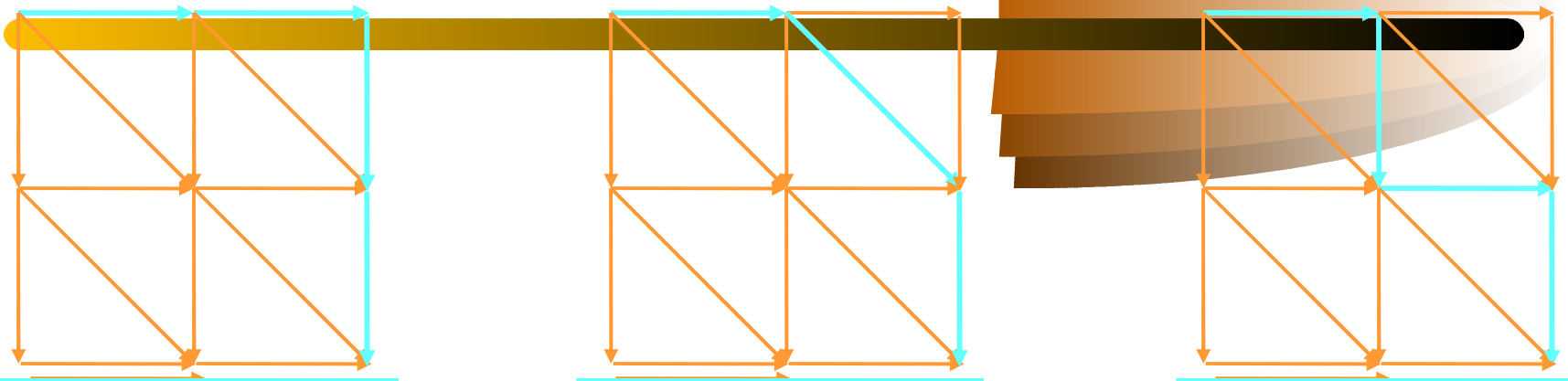
$x_1 x_2 x_3 x_4 x_5$

$y_1 y_2 y_3 y_4 -$

$x_1 \; - \; x_2 x_3 x_4 x_5$

$y_1 y_2 y_3 \; - \; - \; y_4$

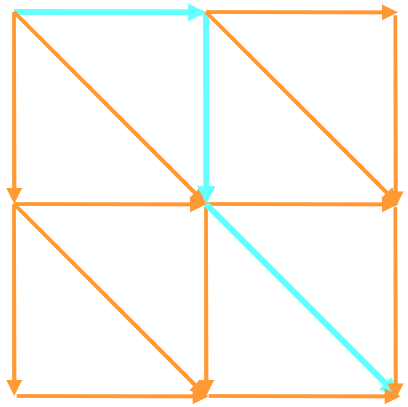
Pfade in F



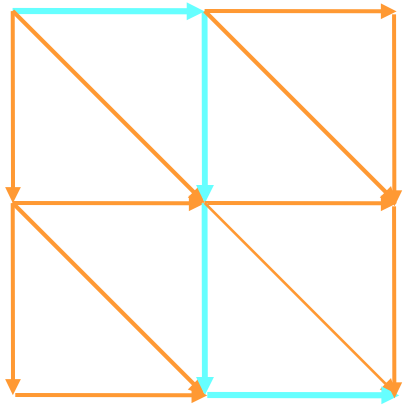
$x_1 x_2 - \quad -$
$- - \quad y_1 y_2$

$x_1 x_2 -$
$- \quad y_1 y_2$

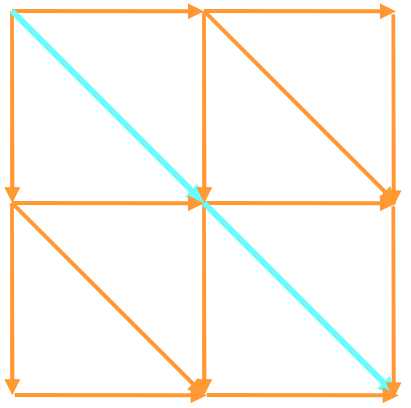
$x_1 - \quad x_2 -$
$- \quad y_1 - \quad y_2$



$x_1 - \quad x_2$
$- \quad y_1 y_2$

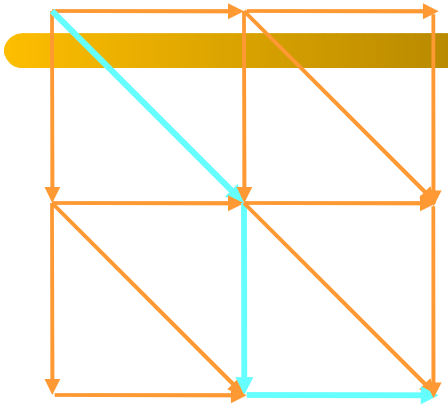


$x_1 - \quad - \quad x_2$
$- \quad y_1 y_2 -$



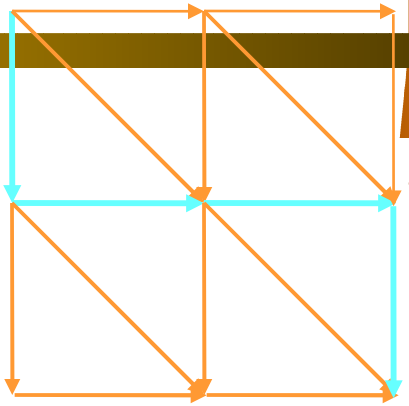
$x_1 x_2$
$y_1 y_2$

Pfade in F



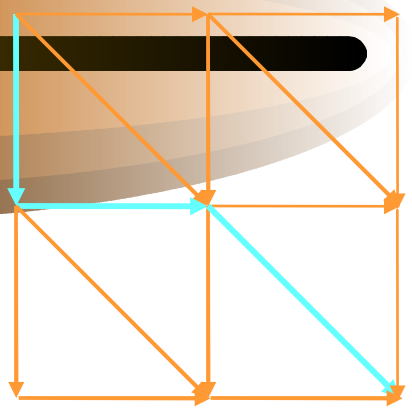
$x_1 \; - \; x_2$

$y_1 y_2 -$



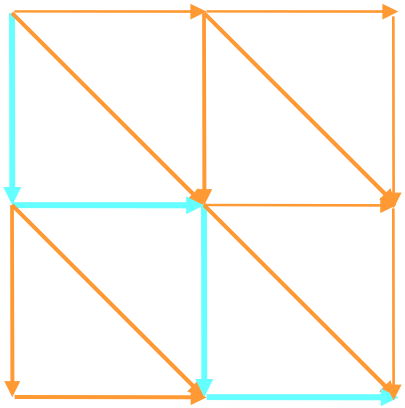
$- \; x_1 x_2 -$

$y_1 - \; - \; y_2$



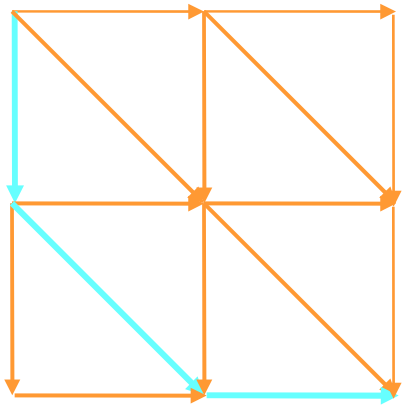
$- \; x_1 x_2$

$y_1 - \; y_2$



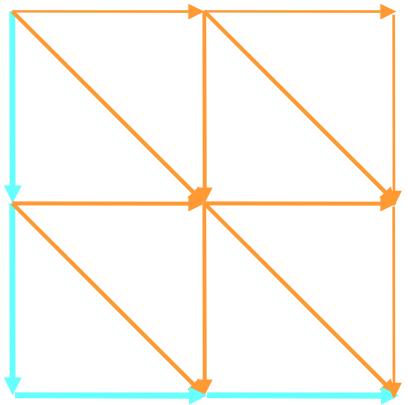
$- \; x_1 - \; x_2$

$y_1 - \; y_2 -$



$- \; x_1 x_2$

$y_1 y_2 -$



$- \; - \; x_1 x_2$

$y_1 y_2 - \; -$

Globales Alignment

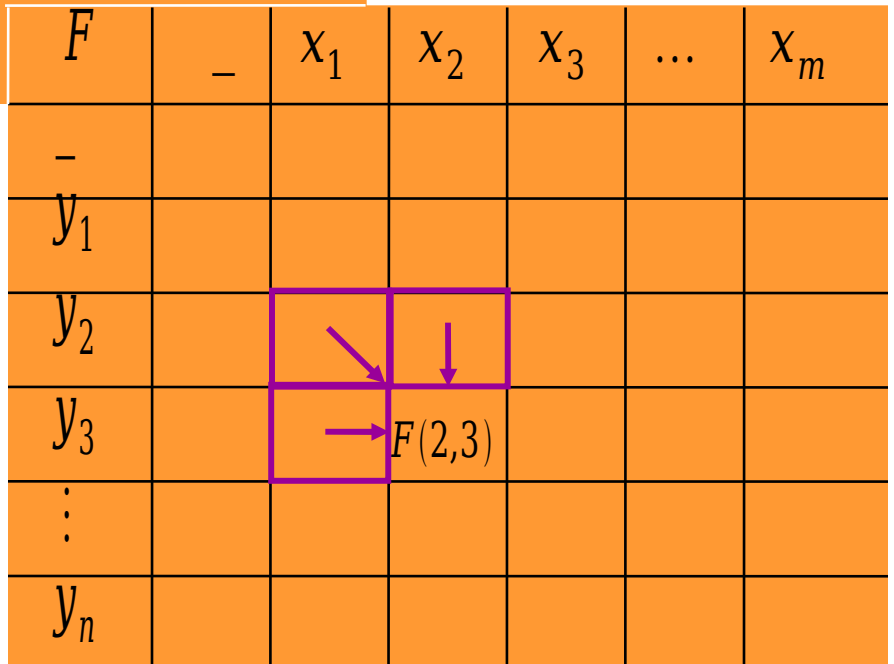
- Needleman-Wunsch Algorithmus
  - benutzt Prinzip der dynamischen Programmierung: optimales Alignment für zwei Sequenzen wird aus optimalen Alignments von Teilsequenzen bestimmt
  - kleinste Einheit: Alignment von zwei Buchstaben (Aminosäuren) bzw. Wert für eine Gap
  - 1. Schritt: Berechnung einer Matrix, die alle möglichen Alignments der Sequenzen repräsentiert. Mit Ausnahme der Initialwerte werden alle Einträge der Matrix mit Hilfe der bereits eingetragenen Werte und einer rekursiven Formel abgeleitet.
  - 2. Schritt: "Ablesen" des besten Alignments aus einem Pfad durch die Matrix (Verfolgung des besten Alignments vom Ziel zum Start)

Needleman-Wunsch Algorithmus

Sequenz 1: $x_1 x_2 x_3 \ldots x_m$

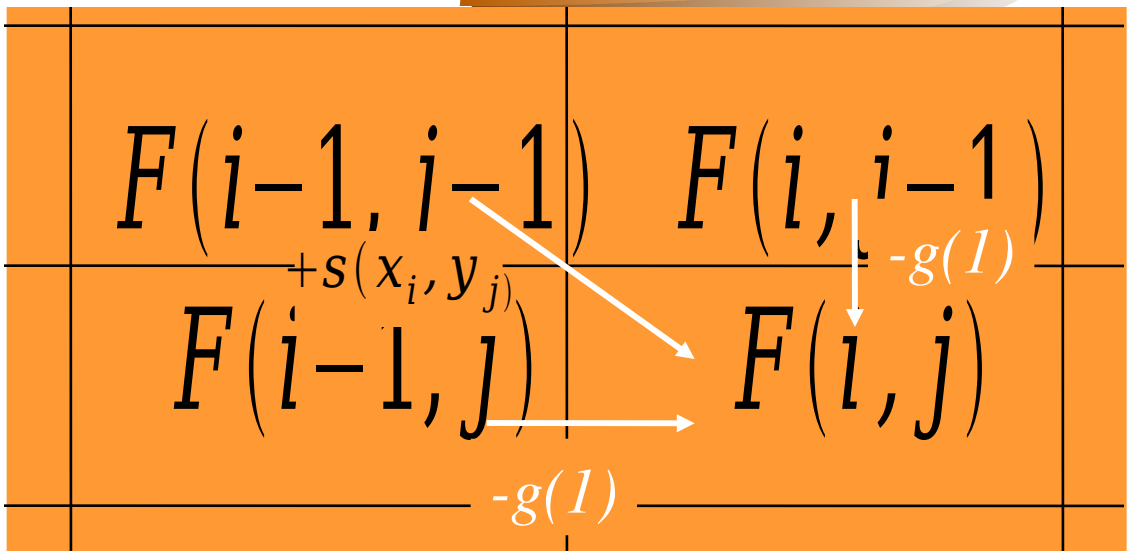
Sequenz 2: $y_1 y_2 y_3 \ldots y_n$

Matrix F wobei $F(i,j)$ den Score für das optimale Alignment der Sequenz $x_1 x_2 \ldots x_i$ mit der Sequenz $y_1 y_2 \ldots y_j$ angibt



Needleman-Wunsch Algorithmus

Zerlegungsprinzip:



$$F(i,j)=\max\begin{cases} F(i-1,j-1)+s(x_i,y_j) \\ F(i-1,j)-g(1) \\ F(i,j-1)-g(1) \end{cases}$$

Needleman-Wunsch: Beispiel

R	D	I	S	L	V

R	N	I	L	V

- PAM 250 (* 10)
- Lineare Gap Penalty mit $q = -6$

Needleman-Wunsch: Beispiel

- Ergebnis-Score: 17
- Alignment

R	D	I	S	L	V
R	N	I	–	L	V

Multiples Alignment

- Vergleich von mehreren (> 2) Sequenzen
- Motivation: Bessere Erkennung von Motiven
  - weniger gut erhaltene Motive gehen im paarweisen Vergleich unter
  - unwichtige Bereiche werden besser ausgeblendet
- Basis für Strukturvorhersage
- Zwei wesentliche Ansätze:
  - simultanes multiples Alignment
  - iteriertes multiples Alignment

Beispiel

8 Fragmente aus Immunoglobulin Sequenzen

```
VTISCTGSSSNIGAG-NHVKWYQQLPG
 VTISCTGTSSNIGS--ITVNWYQQLPG
 LRLSCSSSGFIFSS--YAMYWVRQAPG
 LSLTCTVSGTSFDD--YYSTWVRQPPG
 PEVTCVVVDVSHEDPQVKFNWYVDG--
 ATLVCLISDFYPGA--VTVAWKADS--
 AALGCLVKDYFPEP--VTVSWNSG---
 VSLTCLVKGFYPSD--IAVEWESNG--
```

Konservierte Residuen: W, C
Konservierte Regionen: Q.PG
Aufallende Muster: Hydrophobe Residuen: V,L,P,A,I