Was ist Bioinformatik?

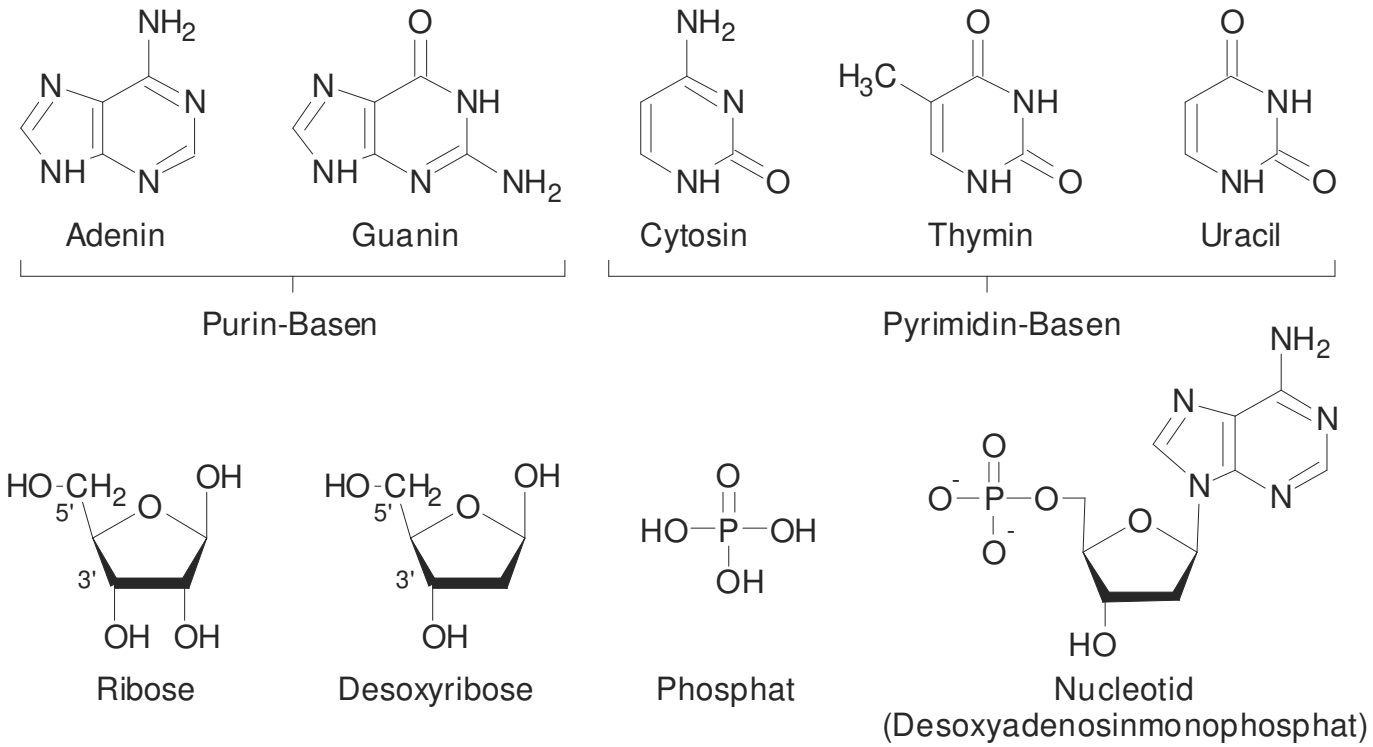
Alexander Vilbig

23. Oktober, 2001

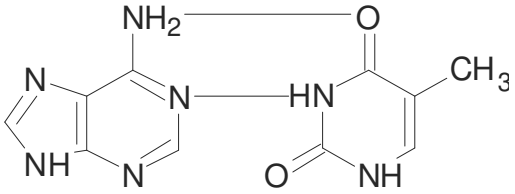
Übersicht

- Biologische Grundlagen
  - DNA, RNA, Gen, Protein, ...
- Informatik in der Biologie
  - Sequenz-Alignment
  - Molekulardynamik-Simulation
- Biologie in der Informatik
  - Evolutionäre Heuristiken
- Zusammenfassung & Ausblick
- Literatur und Ressourcen

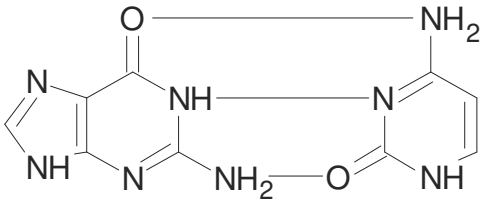
Nukleinsäuren I



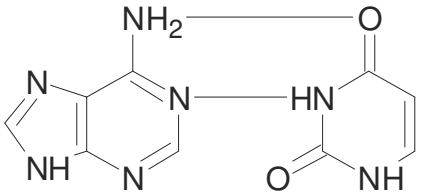
Nukleinsäuren II



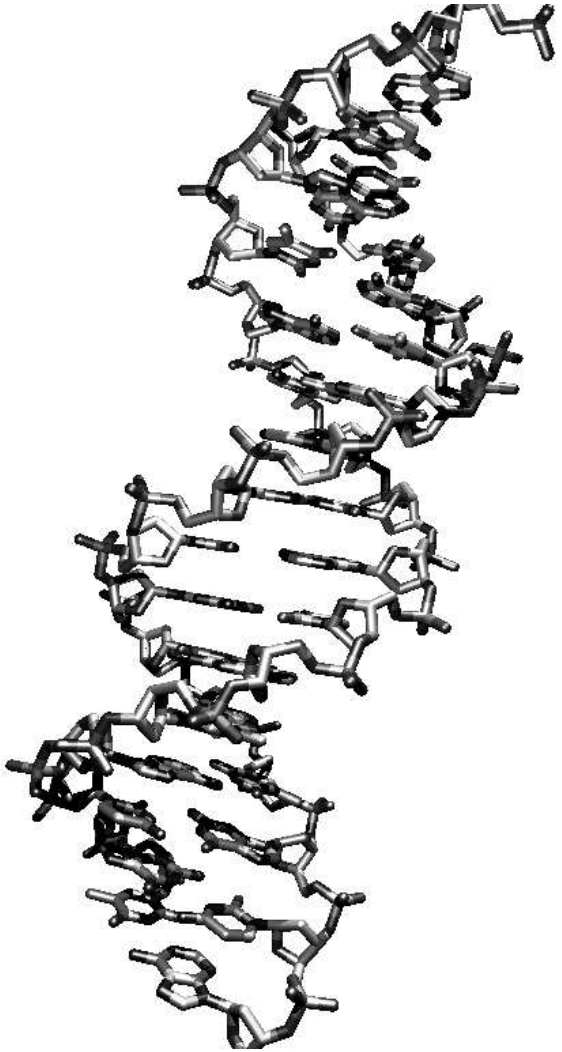
Adenin Thymin



Guanin Cytosin

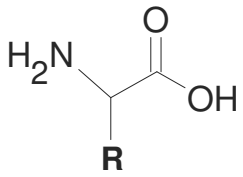


Adenin Uracil

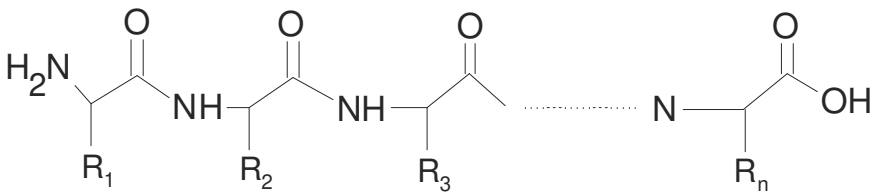


Proteine

Proteine

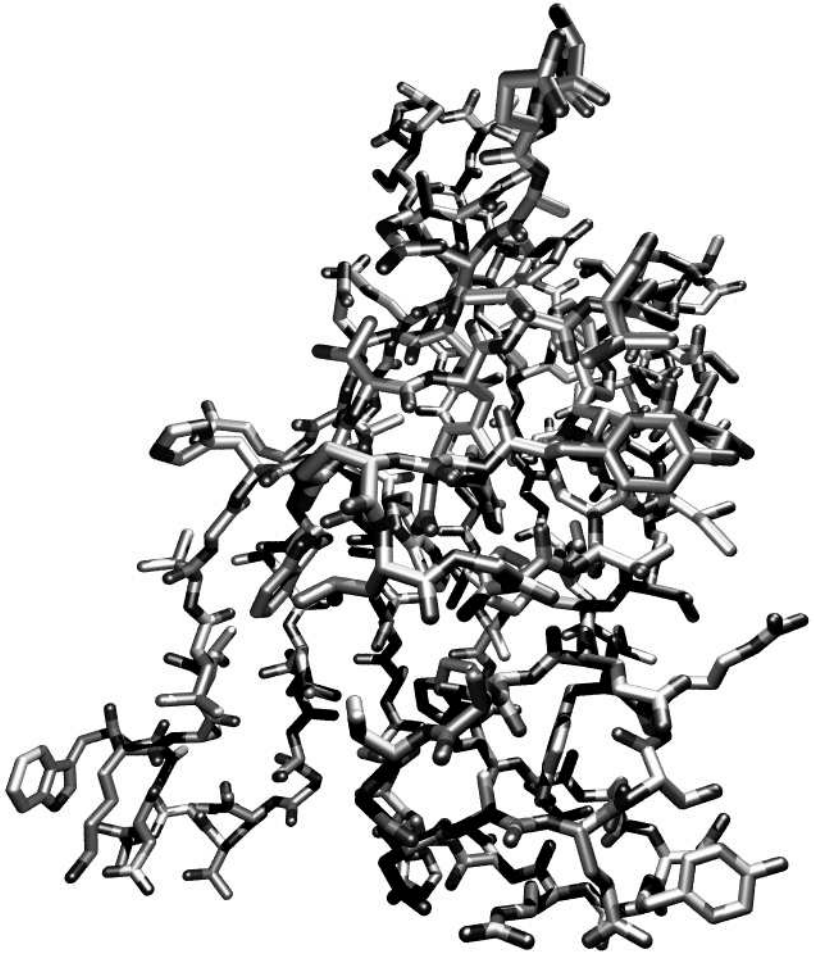


Aminosäure

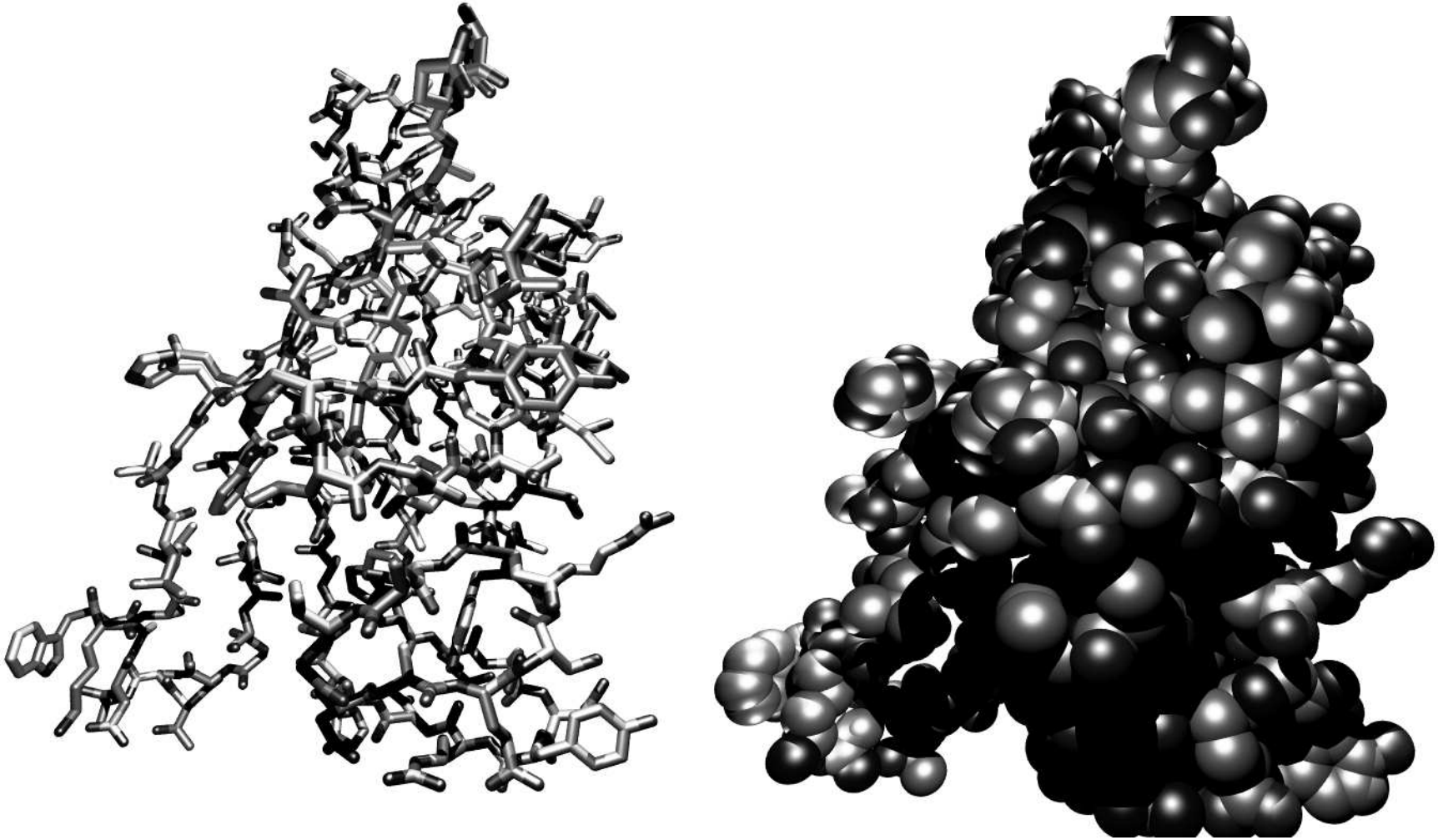


Polypeptid

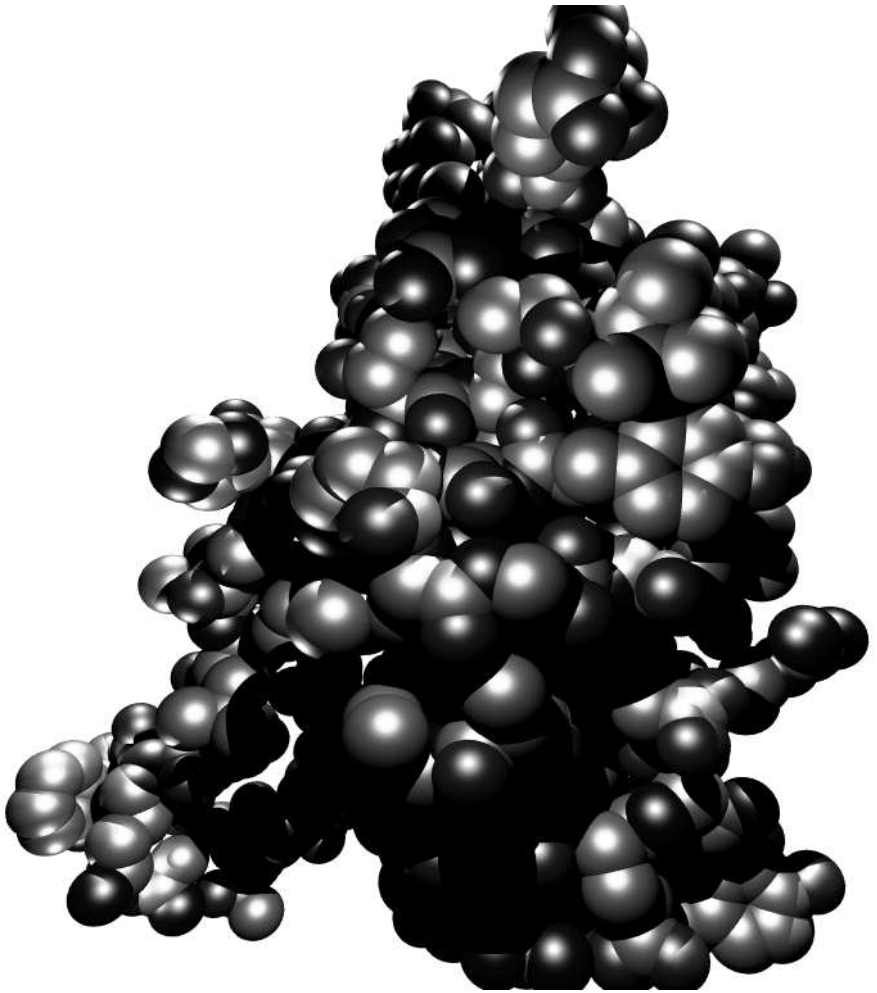
Proteine



Proteine

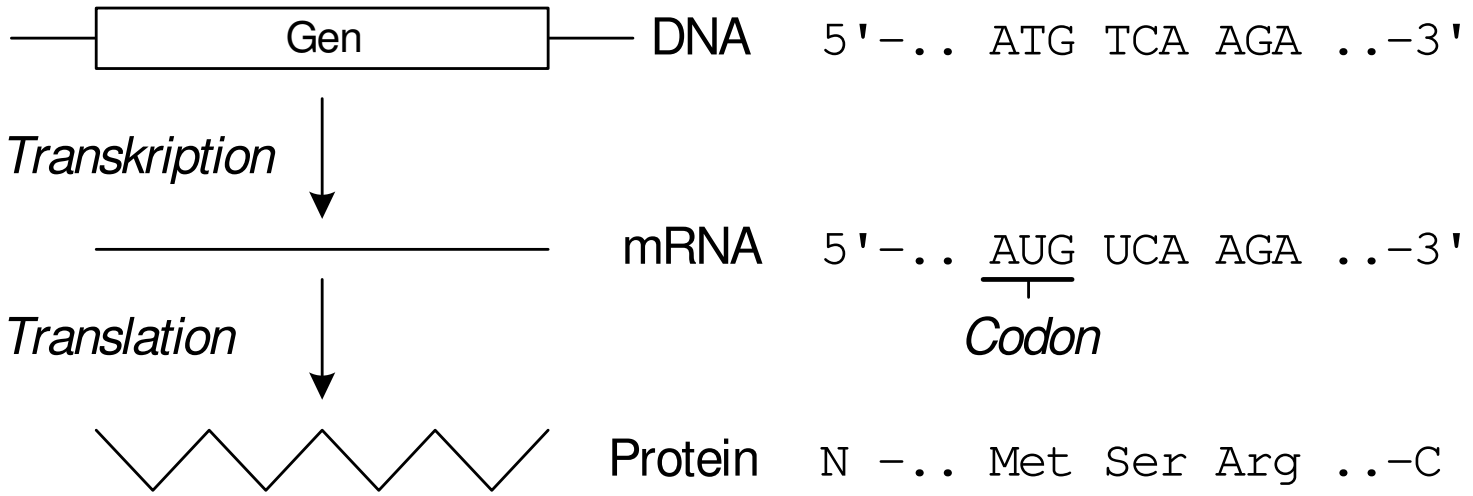


Proteine



Proteine

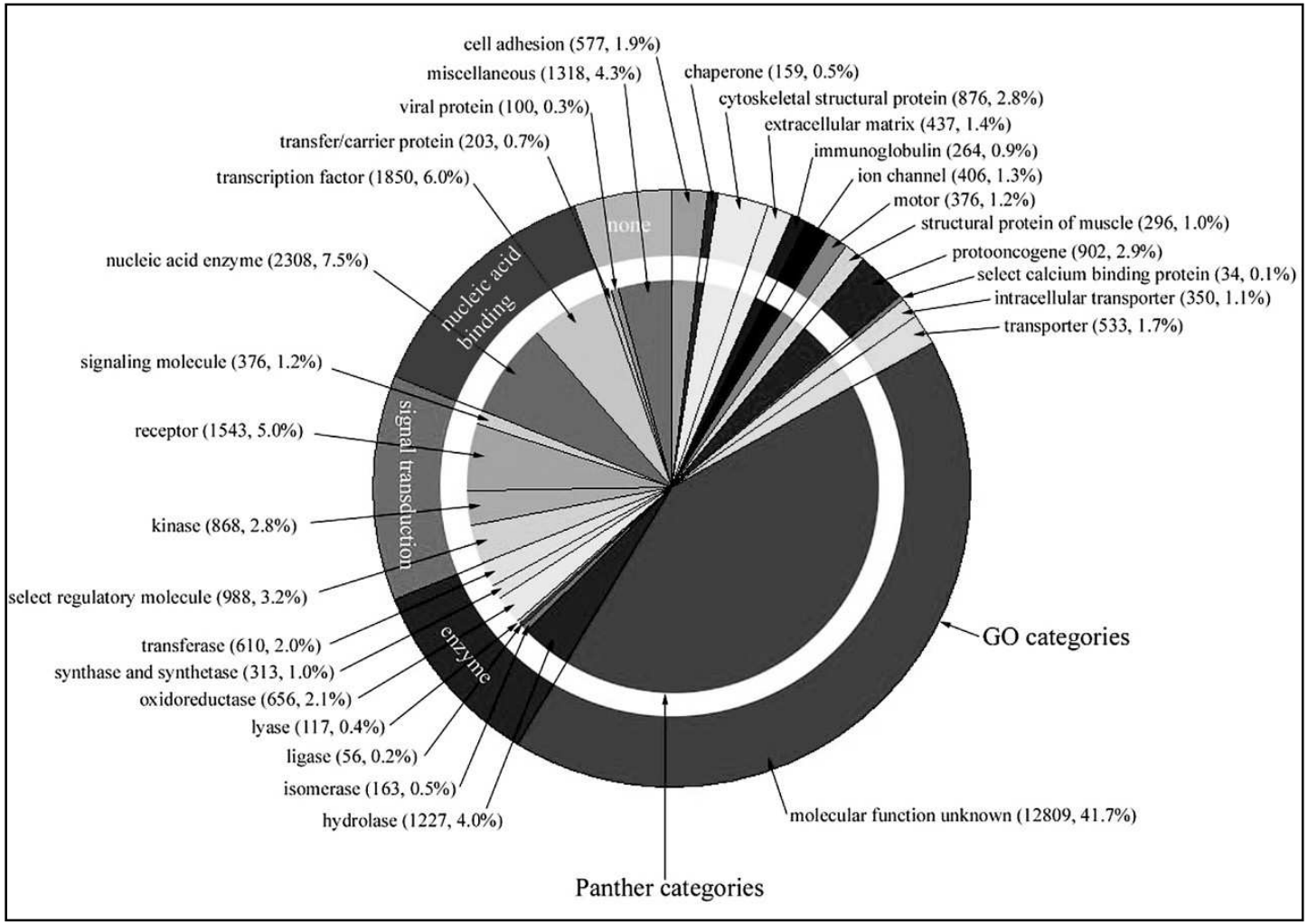
Expression

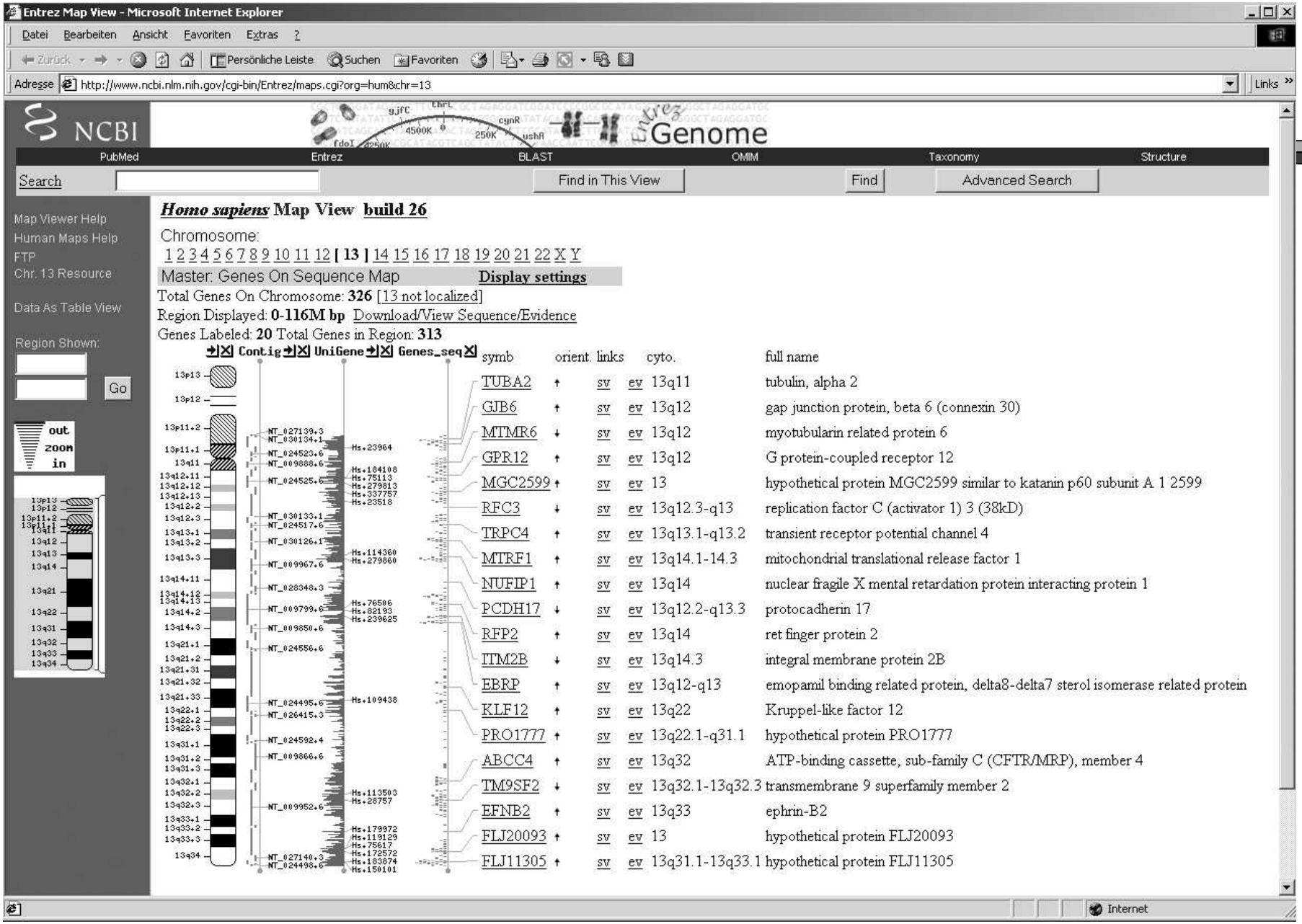

Gen
DNA
5'-.. ATG TCA AGA ..-3'
Transkription
mRNA
5'-.. AUG UCA AGA ..-3'
Translation
Codon
Protein
N -.. Met Ser Arg ..-C

Das menschliche Genom

- Umfaßt ca. 3,2 * 10^9 Basenpaare
- Geschätzte 2 - 4 * 10^4 Gene
- Physisch verteilt auf 23 Chromosomen
- Sequenzierung in 2000/2001 praktisch abgeschlossen
  - Human Genome Project (öffentlich)
  - Celera Genomics Inc. (kommerziell)

Das menschliche Genom




Entrez Map View - Microsoft Internet Explorer
Datei Bearbeiten Ansicht Favoriten Extras ?
Persönliche Leiste Suchen Favoriten
Adresse http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/maps.cgi?org=hum&chr=13
Links
NCBI
Entrez Genome
PubMed Entrez BLAST OMIM Taxonomy Structure
Search
Find in This View
Find
Advanced Search
Map Viewer Help
Human Maps Help
FTP
Chr. 13 Resource
Data As Table View
Region Shown:
Go
out
zoom
in
Homo sapiens Map View build 26
Chromosome:
1 2 3 4 5 6 7 8 9 10 11 12 [13] 14 15 16 17 18 19 20 21 22 X Y
Master: Genes On Sequence Map
Display settings
Total Genes On Chromosome: 326 [13 not localized]
Region Displayed: 0-116M bp Download/View Sequence/Evidence
Genes Labeled: 20 Total Genes in Region: 313
Contig UniGene Genes_seq
symb orient. links cyto. full name
TUBA2 + sv ev 13q11 tubulin, alpha 2
GJB6 + sv ev 13q12 gap junction protein, beta 6 (connexin 30)
MTMR6 - sv ev 13q12 myotubularin related protein 6
GPR12 + sv ev 13q12 G protein-coupled receptor 12
MGC2599 + sv ev 13 hypothetical protein MGC2599 similar to katanin p60 subunit A 1 2599
RFC3 - sv ev 13q12.3-q13 replication factor C (activator 1) 3 (38kD)
TRPC4 + sv ev 13q13.1-q13.2 transient receptor potential channel 4
MTRF1 + sv ev 13q14.1-14.3 mitochondrial translational release factor 1
NUFIP1 + sv ev 13q14 nuclear fragile X mental retardation protein interacting protein 1
PCDH17 - sv ev 13q12.2-q13.3 protocadherin 17
RFP2 + sv ev 13q14 ret finger protein 2
ITM2B - sv ev 13q14.3 integral membrane protein 2B
EBRP + sv ev 13q12-q13 emopamil binding related protein, delta8-delta7 sterol isomerase related protein
KLF12 + sv ev 13q22 Kruppel-like factor 12
PRO1777 + sv ev 13q22.1-q31.1 hypothetical protein PRO1777
ABCC4 + sv ev 13q32 ATP-binding cassette, sub-family C (CFTR/MRP), member 4
TM9SF2 - sv ev 13q32.1-13q32.3 transmembrane 9 superfamily member 2
EFNB2 + sv ev 13q33 ephrin-B2
FLJ20093 + sv ev 13 hypothetical protein FLJ20093
FLJ11305 + sv ev 13q31.1-13q33.1 hypothetical protein FLJ11305
Internet

Informatik in der Biologie

- Datenerfassung und -verwaltung
  - Steuerung von Laborgeräten
  - Labormanagement, Sequence Assembly, ...
  - Datenbanken, Queries, Präsentation, WWW-Anbindung, ...
- Datenanalyse
  - Sequenzbasiert: Alignment, Phylogenie, ...
  - Strukturbasiert: Molecular Docking, Structure Prediction, ...
  - Visualisierung
- Simulation theoretischer Modelle
  - Biochemische Reaktionen
  - Metabolismus
  - Ontogenese, ...

Anwendungsbeispiel 1: Sequenz-Alignment

- „Bestmögliche“ Ausrichtung biomolekularer Sequenzen
  - Insertion, Deletion, Substitution, Match

A = TTAAGCACTTGGATTCTCGGCAGCGTGG
B = TTAAAGGATTCTCCGTAG

Anwendungsbeispiel 1: Sequenz-Alignment

- „Bestmögliche“ Ausrichtung biomolekularer Sequenzen
  - Insertion, Deletion, Substitution, Match

```
A = TTAAGCACTTGGATTCTCGGCAGCGTGG
B = TTAAAGGATTCTCCGTAG
```

Global

```
A´= TTAA-GCACTTGGATTCTCGGCAGCGTGG
B´= TTAAA------GGATTCT---C--CGTAG
```

Anwendungsbeispiel 1: Sequenz-Alignment

- „Bestmögliche" Ausrichtung biomolekularer Sequenzen
  - Insertion, Deletion, Substitution, Match

```
A = TTAAGCACTTGGATTCTCGGCAGCGTGG
B = TTAAAGGATTCTCCGTAG
```

Global	A´= TTAA-GCACTTGGATTCTCGGCAGCGTGG B´= TTAAA------GGATTCT---C--CGTAG
Semi-global	A´= TTAAGCACTTTT---GGATTCTCGGCAG B´= ----------TTAAAGGATTCTCCGTAG

Anwendungsbeispiel 1: Sequenz-Alignment

- „Bestmögliche" Ausrichtung biomolekularer Sequenzen
  - Insertion, Deletion, Substitution, Match

```
A = TTAAGCACTTGGATTCTCGGCAGCGTGG
B = TTAAAGGATTCTCCGTAG
```

Global	A´= TTAA-GCACTTGGATTCTCGGCAGCGTGG B´= TTAAA------GGATTCT---C--CGTAG
Semi-global	A´= TTAAGCACTTTT---GGATTCTCGGCAG B´= ----------TTAAAGGATTCTCCGTAG
mit Gaps	A´= TTAA-GCACTTGGATTCTCGGCAGCGTGG B´= TTAAAG------GATTCTC-----CGTAG

Berechnung eines globalen Alignments I

Distanz $$d(A,B) = \min \sum_{i=1}^{k} d(a_i', b_i') \qquad d(a_i, b_j) = \begin{cases} 0, a_i = b_j \\ > 0, \text{sonst} \end{cases}$$

$$D_{i,j} = \min \begin{pmatrix} D_{i-1,j} + d(a_i, -), \\ D_{i-1,j-1} + d(a_i, b_j), \\ D_{i,j-1} + d(-, b_j) \end{pmatrix}$$

$$D_{i-1,j} = \min \begin{pmatrix} D_{i-2,j} + d(a_{i-1}, -), \\ D_{i-2,j-1} + d(a_{i-1}, b_j), \\ D_{i-1,j-1} + d(-, b_j) \end{pmatrix}$$

Berechnung eines globalen Alignments I

Distanz $$d(A,B) = \min \sum_{i=1}^{k} d(a_i', b_i')$$

$$d(a_i, b_j) = \begin{cases} 0, a_i = b_j \\ > 0, \text{sonst} \end{cases}$$

$$D_{i,j} = \min \begin{pmatrix} D_{i-1,j} + d(a_i, -), \\ D_{i-1,j-1} + d(a_i, b_j), \\ D_{i,j-1} + d(-, b_j) \end{pmatrix}$$

$$D_{i-1,j} = \min \begin{pmatrix} D_{i-2,j} + d(a_{i-1}, -), \\ D_{i-2,j-1} + d(a_{i-1}, b_j), \\ D_{i-1,j-1} + d(-, b_j) \end{pmatrix}$$

Berechnung eines globalen Alignments I

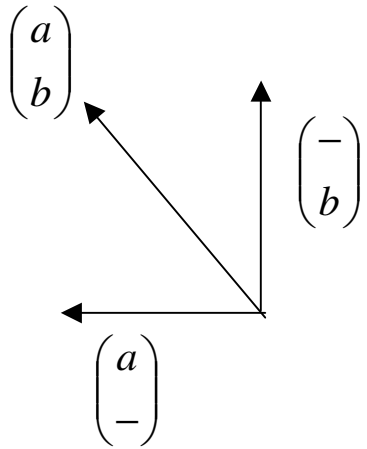
Distanz $d(A,B) = \min \sum_{i=1}^{k} d(a_i', b_i')$ $d(a_i, b_j) = \begin{cases} 0, a_i = b_j \\ > 0, \text{sonst} \end{cases}$

$$D_{i,j} = \min \begin{pmatrix} D_{i-1,j} + d(a_i, -), \\ D_{i-1,j-1} + d(a_i, b_j), \\ D_{i,j-1} + d(-, b_j) \end{pmatrix}$$

$$D_{i-1,j} = \min \begin{pmatrix} D_{i-2,j} + d(a_{i-1}, -), \\ D_{i-2,j-1} + d(a_{i-1}, b_j), \\ D_{i-1,j-1} + d(-, b_j) \end{pmatrix}$$

Berechnung eines globalen Alignments II

		S	O	N	N	E
	0	1	2	3	4	5
M	1	1	2	3	4	5
O	2	2	1	2	3	4
N	3	3	2	1	2	3
D	4	4	3	2	2	3

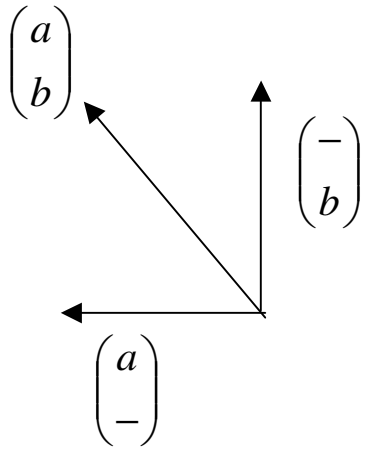


s o n n e

m o - n d

Berechnung eines globalen Alignments II

		S	O	N	N	E
	0	1	2	3	4	5
M	1	1	2	3	4	5
O	2	2	1	2	3	4
N	3	3	2	1	2	3
D	4	4	3	2	2	3

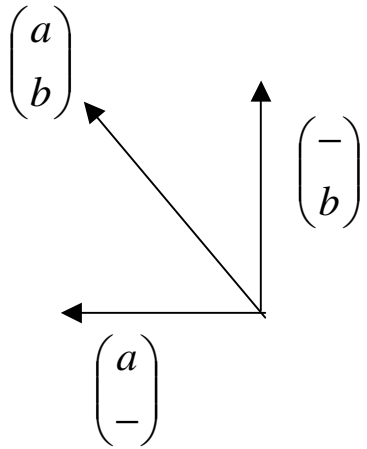


s o n n e

m o - n d

Berechnung eines globalen Alignments II

		S	O	N	N	E
	0	1	2	3	4	5
M	1	1	2	3	4	5
O	2	2	1	2	3	4
N	3	3	2	1	2	3
D	4	4	3	2	2	3

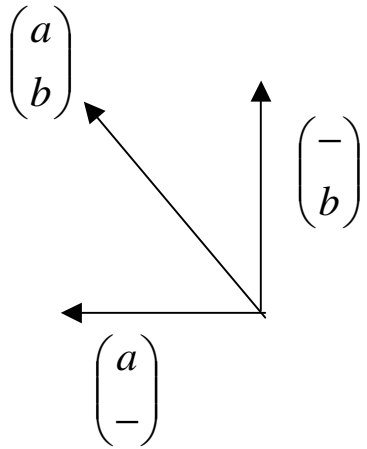


s o n n e

m o - n d

Berechnung eines globalen Alignments II

		S	O	N	N	E
	0	1	2	3	4	5
M	1	1	2	3	4	5
O	2	2	1	2	3	4
N	3	3	2	1	2	3
D	4	4	3	2	2	3




```
s o n n e      sonne
m o - n d      mond-
```

Anwendungsbeispiel 2: Molekulardynamik-Simulation

- Simulation des Verhaltens molekularer Systeme
  - Detaillierungsgrad: individuelle Atome oder funktionale Gruppen
  - Systemgröße: 10^1 - 10^4 Atome + Lösungsmittelatome
  - Zeiträume: 10^{-12} - 10^{-9} s (Pico- bzw. Nanosekunden)
- Zielsetzung
  - Untersuchung von Bindungsvorgängen
  - Berechnung thermodynamischer Größen
  - Drug Design, ...
- Theoretisches Modell
  - Klassische Mechanik vs. Quantenmechanik
  - Newton'sche Bewegungsgleichungen für große Systeme

Molekulardynamik-Simulation I

- Algorithmus von Verlet
  - Integration der Bewegungsgleichungen
  - Kombination der Taylor-Entwicklung an den Stellen $t+\Delta t$ und $t-\Delta t$

$$\frac{\partial \boldsymbol{r}_i(t)}{\partial t} = \boldsymbol{v}_i(t) \qquad \frac{\partial \boldsymbol{v}_i(t)}{\partial t} = \frac{\boldsymbol{F}_i(t)}{m_i}$$

Molekulardynamik-Simulation I

- Algorithmus von Verlet
  - Integration der Bewegungsgleichungen
  - Kombination der Taylor-Entwicklung an den Stellen $t+\Delta t$ und $t-\Delta t$

$$\frac{\partial \boldsymbol{r}_i(t)}{\partial t} = \boldsymbol{v}_i(t) \qquad \frac{\partial \boldsymbol{v}_i(t)}{\partial t} = \frac{\boldsymbol{F}_i(t)}{m_i}$$

$$\boldsymbol{r}_i(t+\Delta t) = 2\,\boldsymbol{r}_i(t) - \boldsymbol{r}_i(t-\Delta t) + \frac{\boldsymbol{F}_i(t)}{m_i}\Delta t^2 + O\left(\Delta t^4\right)$$

Molekulardynamik-Simulation I

- Algorithmus von Verlet
  - Integration der Bewegungsgleichungen
  - Kombination der Taylor-Entwicklung an den Stellen $t+\Delta t$ und $t-\Delta t$

$$\frac{\partial \boldsymbol{r}_i(t)}{\partial t} = \boldsymbol{v}_i(t) \qquad \frac{\partial \boldsymbol{v}_i(t)}{\partial t} = \frac{\boldsymbol{F}_i(t)}{m_i}$$

$$\boldsymbol{r}_i(t+\Delta t) = 2\,\boldsymbol{r}_i(t) - \boldsymbol{r}_i(t-\Delta t) + \frac{\boldsymbol{F}_i(t)}{m_i}\Delta t^2 + O\left(\Delta t^4\right)$$

$$\boldsymbol{F}_i(t) = -\frac{\partial U\left[\boldsymbol{r}_1(t), \boldsymbol{r}_2(t), \ldots, \boldsymbol{r}_N(t)\right]}{\partial \boldsymbol{r}_i(t)}$$

Molekulardynamik-Simulation I

- Algorithmus von Verlet
  - Integration der Bewegungsgleichungen
  - Kombination der Taylor-Entwicklung an den Stellen t+Δt und t-Δt

$$\frac{\partial \boldsymbol{r}_i(t)}{\partial t} = \boldsymbol{v}_i(t) \qquad \frac{\partial \boldsymbol{v}_i(t)}{\partial t} = \frac{\boldsymbol{F}_i(t)}{m_i}$$

$$\boldsymbol{r}_i(t+\Delta t) = 2\boldsymbol{r}_i(t) - \boldsymbol{r}_i(t-\Delta t) + \frac{\boldsymbol{F}_i(t)}{m_i}\Delta t^2 + O\left(\Delta t^4\right)$$

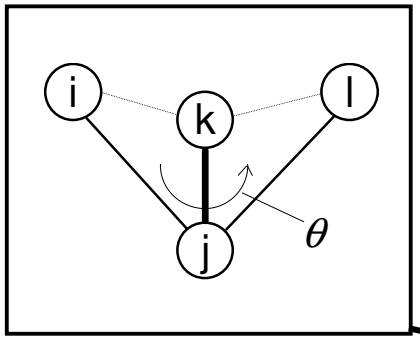
$$\boldsymbol{F}_i(t) = -\frac{\partial U\left[\boldsymbol{r}_1(t), \boldsymbol{r}_2(t), \ldots, \boldsymbol{r}_N(t)\right]}{\partial \boldsymbol{r}_i(t)}$$

$$U\left(\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N\right) = E_B + E_\theta + E_\varphi + E_\psi + E_{VDW} + E_C$$

Molekulardynamik-Simulation I

- Algorithmus von Verlet
  - Integration der Bewegungsgleichungen
  - Kombination der Taylor-Entwicklung an den Stellen t+Δt und t-Δt

$$\frac{\partial \boldsymbol{r}_i(t)}{\partial t} = \boldsymbol{v}_i(t) \qquad \frac{\partial \boldsymbol{v}_i(t)}{\partial t} = \frac{\boldsymbol{F}_i(t)}{m_i}$$



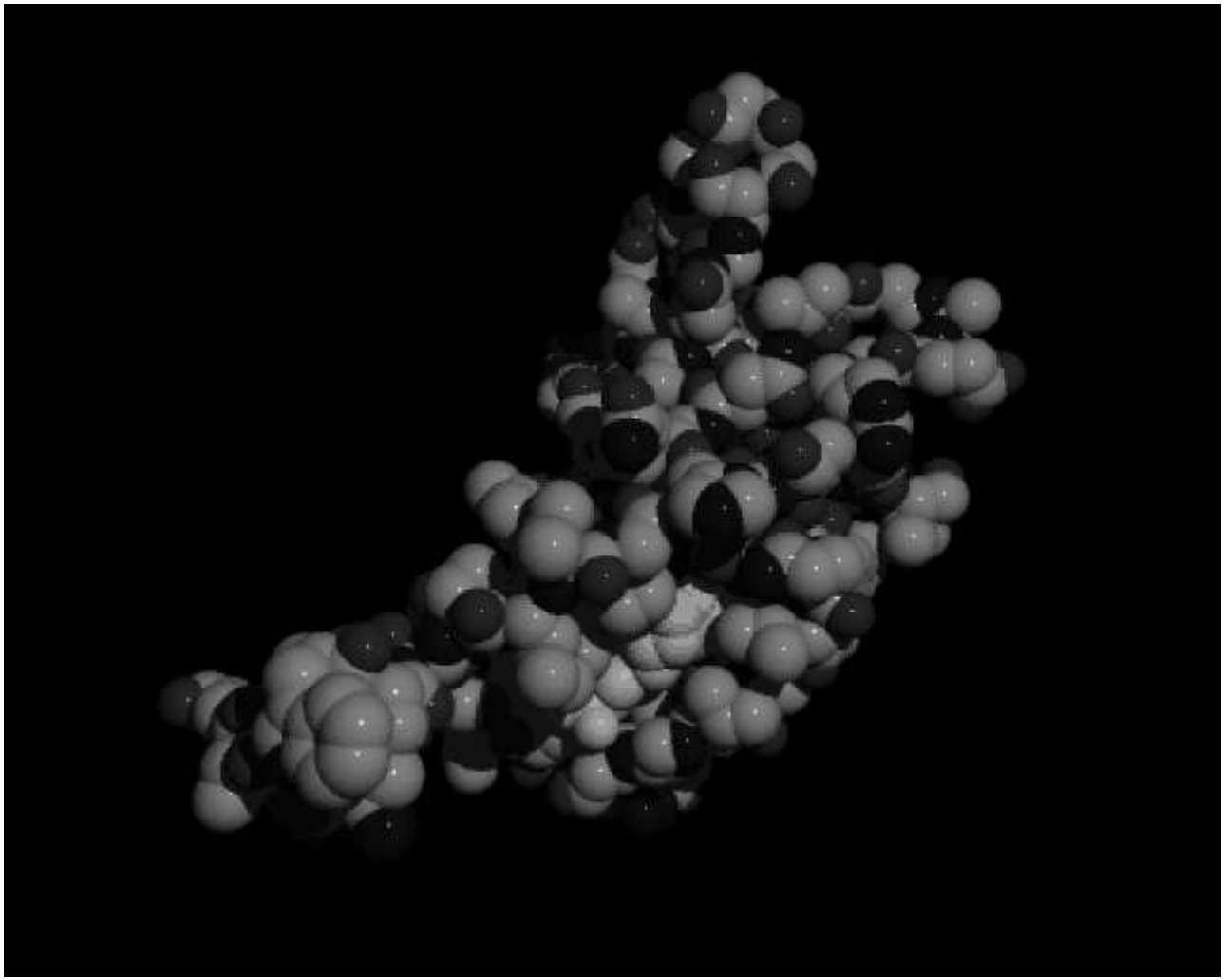
$$\boldsymbol{r}_i(t+\Delta t) = 2\boldsymbol{r}_i(t) - \boldsymbol{r}_i(t-\Delta t) + \frac{\boldsymbol{F}_i(t)}{m_i}\Delta t^2 + O\left(\Delta t^4\right)$$

$$\boldsymbol{F}_i(t) = -\frac{\partial U\left[\boldsymbol{r}_1(t), \boldsymbol{r}_2(t), \ldots, \boldsymbol{r}_N(t)\right]}{\partial \boldsymbol{r}_i(t)}$$

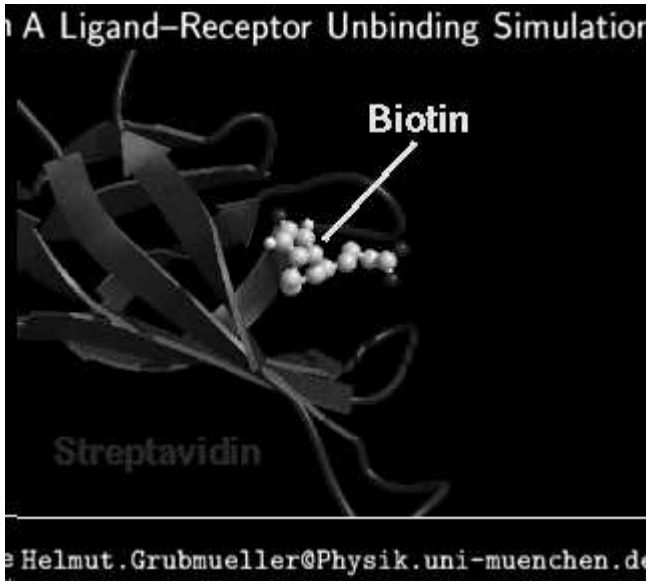
$$U\left(\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N\right) = E_B + E_\theta + E_\varphi + E_\psi + E_{VDW} + E_C$$

Molekulardynamik-Simulation II

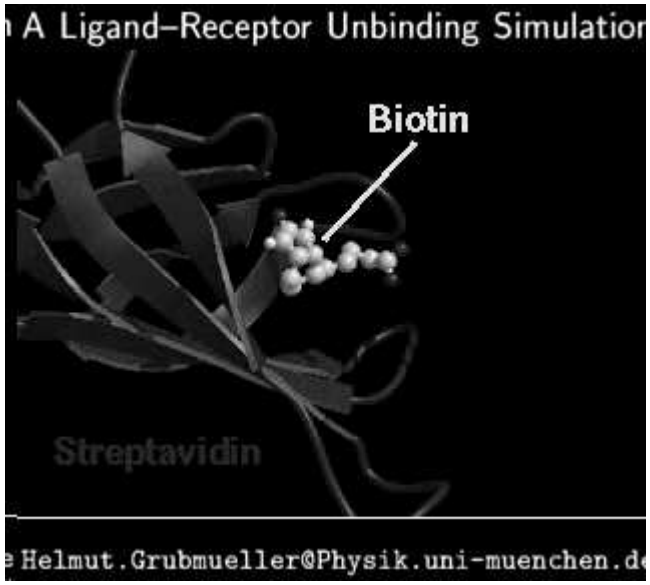
Molekulardynamik-Simulation II

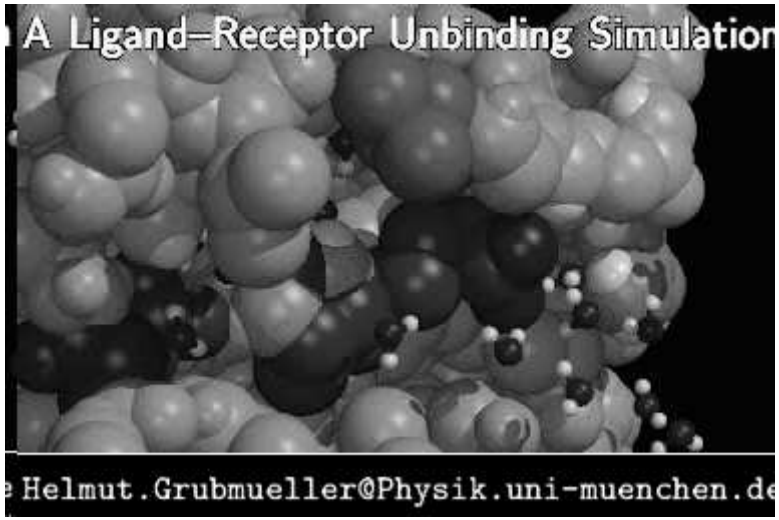


Molekulardynamik-Simulation II



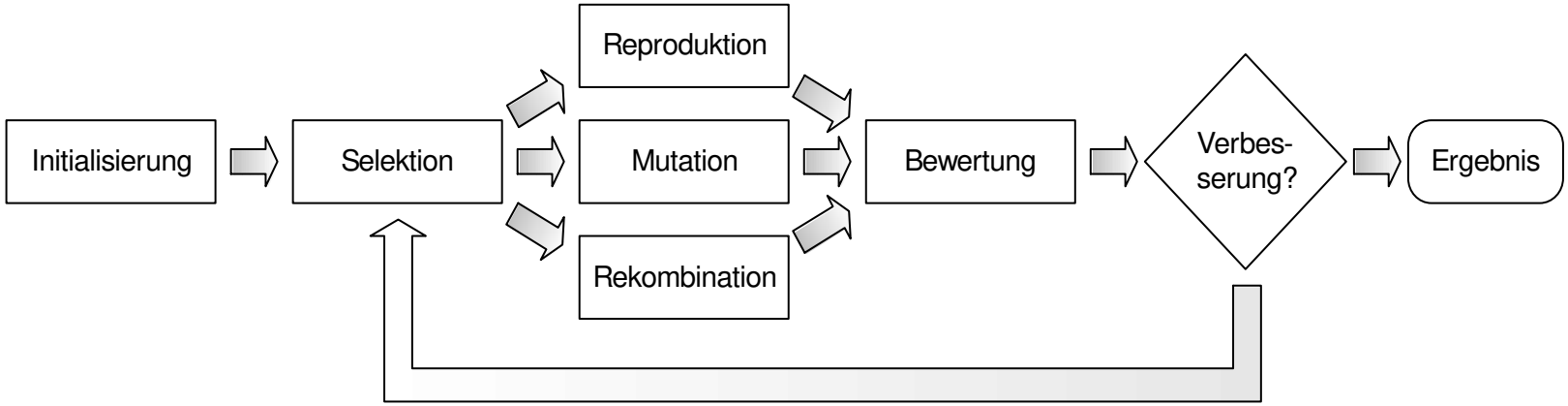
Molekulardynamik-Simulation II



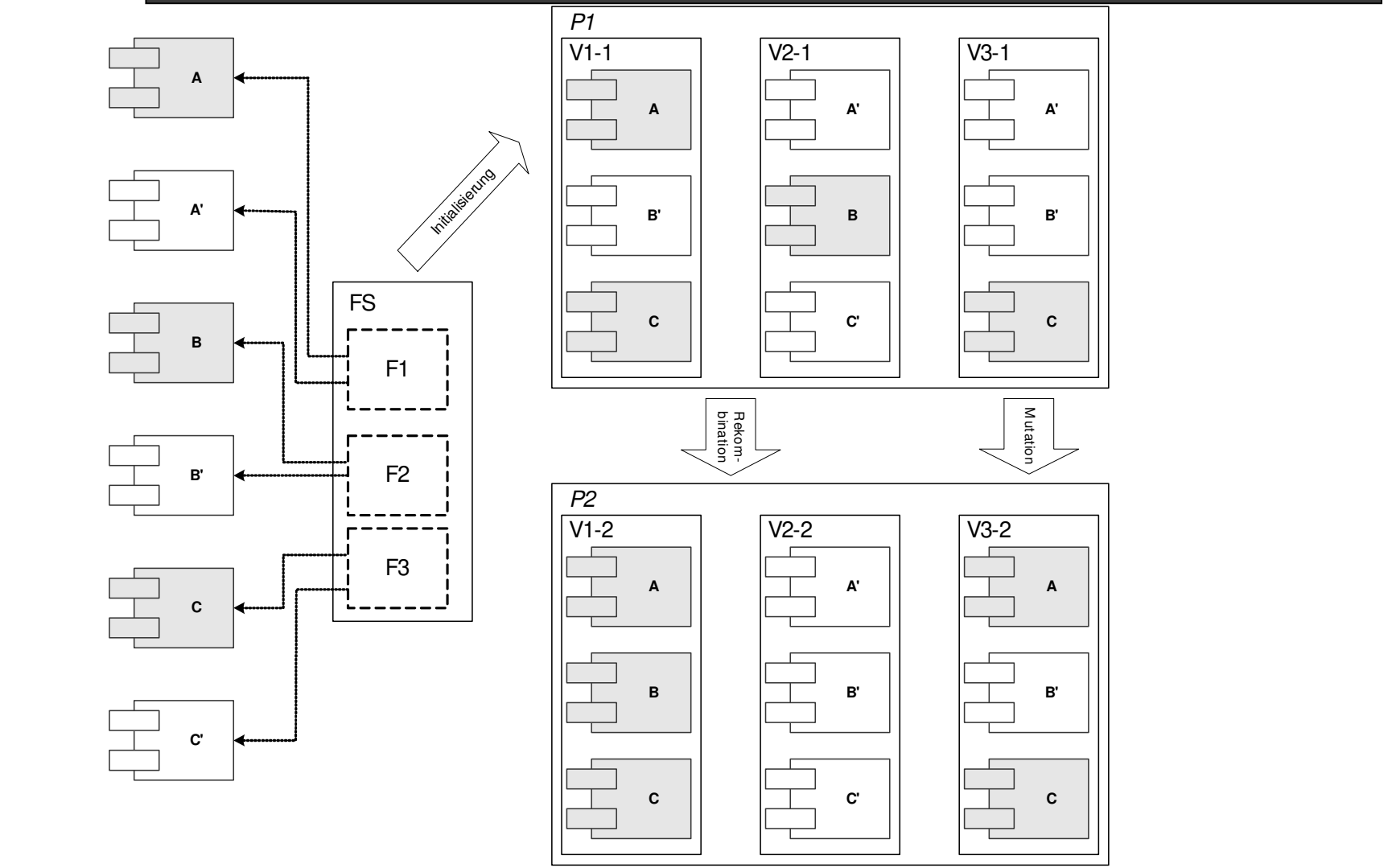


Anwendungsbeispiel 3: Genetische Algorithmen

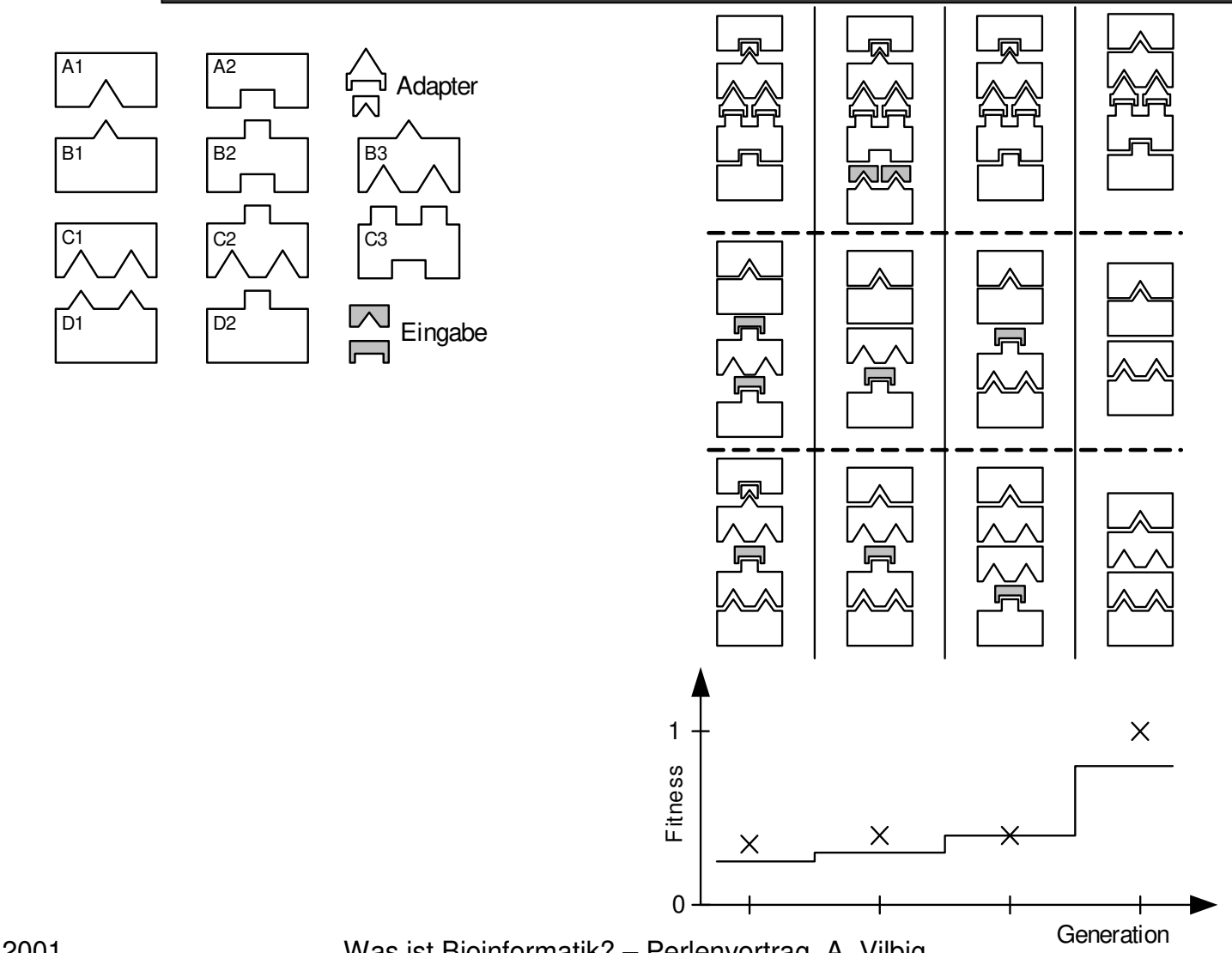
- Evolutionäre Heuristik zur kombinatorischen Optimierung
  - Anwendung bei komplexer, nicht-stetiger Zielfunktion
  - Betrachtung einer Menge von Lösungen als *Individuen* einer *Population*, welche sich durch *Selektion* und fortgesetzte Anwendung der *Genetischen Operatoren* weiterentwickelt
  - Lineare Kodierung der Lösung erforderlich $\leftrightarrow$ *Chromosom*
  - Theorie nur für einfache Kodierung ausgearbeitet („Bitstrings")
  - Anpassung der *Fitneß-Funktion* und Gen. Operatoren erforderlich



Genetische Algorithmen I



Genetische Algorithmen II



Zusammenfassung & Ausblick

- Informatik in der Biologie
  - Essentiell zur Bewältigung großer Datenmengen
  - Anwendung „traditioneller" Algorithmen zur Klärung wichtiger biologischer Fragestellungen
  - Vielfältige Lösungsansätze
- Biologie in der Informatik
  - Evolutionäre Heuristiken als interessante Alternative zu bekannten Verfahren
  - Erfassung und Verarbeitung unscharfer Informationen
  - Potential bei weitem nicht ausgeschöpft
- Zukünftige Fragestellungen & Ansätze
  - Genomics, Proteomics: Holistische Betrachtung komplexer biomolekularer Systeme
  - Protein Folding: Zusammenhang zwischen Sequenz und Struktur
  - Vision: „Organic Software Engineering" - Analogie zwischen Software-System und Organismus?

Literatur

- Biochemie
  - L. Stryer, *Biochemistry.* Freeman & Co, New York 1988.
- Genetik
  - B. Lewin, *Genes IV.* Oxford University Press, New York 1990.
- Sequenzanalyse
  - M. Waterman, *Introduction to Computational Biology: Maps, Sequences and Genomes.* Chapman & Hall, London 1995.
- Molekulardynamik-Simulation
  - J. M. Haile, *Molecular Dynamics Simulation: Elementary Methods.* Wiley-Interscience, New York, 1992.
- Genetische Algorithmen
  - D. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley 1989.

WWW-Ressourcen

- Human Genome Project
  - www.ornl.gov/hgmis/, www.nhgri.nih.gov/HGP/
- Forschungseinrichtungen
  - National Center for Biotechnology Information: www.ncbi.nlm.nih.gov/
  - European Bioinformatics Institute: www.ebi.ac.uk/
  - European Molecular Biology Laboratory: www.embl-heidelberg.de/
- Online-Dokumente und URLs
  - Folding@Home: www.stanford.edu/group/pandegroup/cosm/
  - MD Simulation des Streptavidin-Biotin Komplexes (MPI Göttingen)
    www.mpibpc.gwdg.de/abteilungen/071/streptavidin.html
  - BioComputing Hypertext Coursebook der Universität Bielefeld
    www.techfak.uni-bielefeld.de/bcd/Curric/welcome.html
  - Vorlesung Scientific Computation an der ETH Zürich
    linneus20.ethz.ch:8080/wsrscript.html
  - Lehrstuhl für effiziente Algorithmen der TU München
    wwwmayr.informatik.tu-muenchen.de/