

## Inhaltsverzeichnis

**Vorwort** XV

**Website** XVII

<b>1</b>	<b>Biologische Grundlagen</b>	3
1.1	DNA	3
1.2	Genetischer Code und Genomkomposition	5
1.3	Transkription	9
1.4	RNA	10
1.5	Proteine	11
1.6	Peptidbindung	13
1.7	Konformation von Aminosäureseitenketten	14
1.8	Ramachandran-Plot	15
1.9	Hierarchische Beschreibung von Proteinstrukturen	16
1.10	Sekundärstrukturelemente	16
1.11	$\alpha$ -Helix	17
1.12	$\beta$ -Faltblätter	17
1.13	Supersekundärstrukturelemente	18
1.14	Protein-Domänen	19
1.15	Proteinfamilien	20
1.16	Fachbegriffe	23
1.17	Zitierte Literatur	25
<b>2</b>	<b>Sequenzen und ihre Funktion</b>	27
2.1	Definitionen und Operatoren	28
2.2	DNA-Sequenzen	29
2.3	Proteinsequenzen	29
2.4	Vergleich der Sequenzkomposition	33
2.5	Ontologien	35
2.6	Semantische Ähnlichkeit von GO-Termen	38
2.7	Zitierte Literatur	40

<b>3</b>	<b>Datenbanken</b>	41
3.1	DNA-Sequenz-Datenbanken	42
3.2	RNA-Sequenz-Datenbanken	43
3.3	Proteinsequenz-Datenbanken	44
3.4	Proteinstruktur-Datenbanken	45
3.5	SMART: Analyse der Domänenarchitektur	46
3.6	STRING: Proteine und ihre Interaktionen	47
3.7	SCOP: Strukturelle Klassifikation von Proteinen	47
3.8	Pfam: Kompilation von Proteinfamilien	49
3.9	COG und eggNOG: Gruppen orthologer Gene	50
3.10	Weitere Datenbanken	50
3.11	Zitierte Literatur	54
<b>4</b>	<b>Grundbegriffe der Stochastik</b>	59
4.1	Grundbegriffe der beschreibenden Statistik	61
4.2	Urnenexperimente und diskrete Verteilungen	63
4.3	Die Kolmogoroffschen Axiome	66
4.4	Bedingte Wahrscheinlichkeit und Unabhängigkeit	67
4.5	Zufallselemente	68
4.6	Unabhängigkeit von Zufallselementen	71
4.7	Markov-Ketten	71
4.8	Erwartungswerte	72
4.9	Varianzen	74
4.10	Wichtige Wahrscheinlichkeitsverteilungen	77
4.10.1	Diskrete Verteilungen	78
4.10.2	Totalstetige Verteilungen	79
4.11	Schätzer	82
4.12	Grundlagen statistischer Tests	85
4.13	Eine optimale Entscheidungstheorie: Die Neyman-Pearson-Methode	86
4.14	Zitierte Literatur	87
<b>5</b>	<b>Bayessche Entscheidungstheorie und Klassifikatoren</b>	89
5.1	Bayessche Entscheidungstheorie	89
5.1.1	Ein Beispiel: Klassifikation der Proteinoberfläche	90
5.1.2	Übergang zu bedingten Wahrscheinlichkeiten	91
5.1.3	Erweitern auf $m$ Eigenschaften	93
5.2	Marginalisieren	95
5.3	Boosting	96
5.4	ROC-Kurven	98
5.4.1	Gewichten der Fehlklassifikationen	99
5.4.2	Aufnehmen einer ROC-Kurve	99
5.5	Testmethoden für kleine Trainingsmengen	101
5.6	Zitierte Literatur	104

<b>6</b>	<b>Klassische Cluster- und Klassifikationsverfahren</b>	105
6.1	Metriken und Clusteranalyse	106
6.2	Das mittlere Fehlerquadrat als Gütemaß bei Clusteralgorithmen	106
6.3	Ein einfaches iteratives Clusterverfahren	108
6.4	<i>k-Means</i> -Clusterverfahren	110
6.4.1	Wahl einer geeigneten Anzahl $k$ von Clustern	111
6.4.2	Statistische Bewertung der Clusteranzahl	111
6.5	Hierarchische Clusterverfahren	113
6.6	Nächster-Nachbar-Klassifikation	114
6.7	$k$ nächste Nachbarn	115
6.8	Zitierte Literatur	117
<b>7</b>	<b>Neuronale Netze</b>	119
7.1	Architektur von neuronalen Netzen	120
7.2	Das Perzeptron	121
7.2.1	Schwellenwertfunktion	121
7.2.2	Ein Beispiel: Modellierung Boolscher Funktionen	122
7.3	Lösbarkeit von Klassifikationsaufgaben	123
7.4	Universelle Approximation	126
7.5	Lernen in neuronalen Netzen	128
7.5.1	Der Backpropagation-Algorithmus	129
7.5.2	Interpretation des Lernschrittes	131
7.6	Codierung der Eingabe	132
7.7	Selbstorganisierende Karten	133
7.7.1	Aufbau der Karte	134
7.7.2	Selbstorganisation	135
7.8	Zitierte Literatur	136
<b>8</b>	<b>Genetische Algorithmen</b>	137
8.1	Objekte und Funktionen	139
8.2	Algorithmus	141
8.3	Der Begriff des Schemas	142
8.4	Dynamik der Anzahl von Schemata	143
8.5	Codierung der Problemstellung	145
8.6	Genetisches Programmieren	146
8.7	Zitierte Literatur	149
<b>9</b>	<b>Paarweiser Sequenzvergleich</b>	153
9.1	Dotplots	155
9.1.1	Definition	155
9.1.2	Beispiel	155
9.1.3	Implementierung	157
9.1.4	Abschätzen der Laufzeit	158
9.1.5	Anwendungen	159
9.1.6	Einschränkungen und Ausblick	162

9.2	Entwicklung eines optimalen Alignmentverfahrens	162
9.2.1	Vom paarweisen zum multiplen Sequenzalignment	164
9.2.2	Dynamisches Programmieren	165
9.2.3	Distanz, Metrik	167
9.2.4	Minkowski-Metrik	168
9.2.5	Eine Metrik für Zeichenketten: Die Hamming-Distanz	169
9.3	Levenshtein-Distanz	170
9.3.1	Berechnung der Levenshtein-Distanz	172
9.3.2	Ableiten des Alignments	175
9.4	Bestimmen der Ähnlichkeit von Sequenzen	176
9.4.1	Globales Alignment	177
9.4.2	Lokales Sequenzalignment	177
9.5	Optimales Bewerten von Lücken	179
9.5.1	Bewertung mithilfe affiner Kostenfunktion	180
9.5.2	Integration in Algorithmen	180
9.6	Namensgebung	182
9.7	Zitierte Literatur	182
<b>10</b>	<b>Sequenz-Motive</b>	183
10.1	Signaturen	184
10.2	Die PROSITE-Datenbank	185
10.3	Die BLOCKS-Datenbank	186
10.4	Sequenz-Profile	187
10.5	Bestimmen von Scores für Promotor-Sequenzen	188
10.6	Sequenz-Logos	189
10.7	Konsensus-Sequenzen	189
10.8	Sequenzen niedriger Komplexität	191
10.9	Der SEG-Algorithmus	191
10.10	Zitierte Literatur	195
<b>11</b>	<b>Scoring-Schemata</b>	197
11.1	Zur Theorie von Scoring-Matrizen	198
11.2	Algorithmen bedingte Anforderung an Scoring-Matrizen	200
11.3	Identitätsmatrizen	201
11.4	PAM-Einheit	201
11.5	PAM-Matrizen	202
11.6	Erweiterte Datenbasis: Die JTT-Matrix	203
11.7	BLOSUM-Matrizen	205
11.8	Matrix-Entropie	207
11.9	Scoring-Schemata und Anwendungen	209
11.10	Scoring-Funktionen	209
11.11	Zitierte Literatur	210

<b>12</b>	<b>FASTA, BLAST, PSI-BLAST</b>	213
12.1	FASTA	215
12.2	FASTA-Statistik	217
12.3	BLAST	219
12.4	Statistik von Alignments	222
12.4.1	Statistik globaler Alignments	222
12.4.2	Statistik lokaler Alignments	222
12.5	Vergleich der Empfindlichkeit von FASTA und BLAST	227
12.6	Verfeinerung der Algorithmen	228
12.7	Profil basierter Sequenzvergleich	228
12.8	Verwenden von Intermediärsequenzen	229
12.9	PSI-BLAST	231
12.10	Die Empfindlichkeit von Sequenzvergleichsmethoden	235
12.11	Vergleich von Profilen und Konsensus-Sequenzen	236
12.12	Zitierte Literatur	238
<b>13</b>	<b>Multiple Sequenzalignments</b>	239
13.1	Berechnen von Scores für multiple Sequenzalignments	241
13.2	Iteratives, progressives Bestimmen eines multiplen Alignments	242
13.3	ClustalW: Konzepte	243
13.4	ClustalW: Algorithmus	244
13.5	ClustalW: Multiples Sequenzalignment für Trypsin-Inhibitoren	244
13.6	T-Coffee	246
13.7	M-Coffee und 3D-Coffee	250
13.8	Alternative Ansätze	251
13.9	Verwenden von MSAs zur Charakterisierung von Residuen	251
13.9.1	Entwickeln der Scoring-Funktion	252
13.9.2	SDPpred: Vergleich homologer Proteine mit unterschiedlicher Spezifität	254
13.10	Alignment von DNA- und RNA-Sequenzen	256
13.11	Zitierte Literatur	257
<b>14</b>	<b>Grundlagen phylogenetischer Analysen</b>	259
14.1	Phylogenetische Ansätze	263
14.2	Distanz basierte Verfahren	264
14.2.1	Ultrametrische Matrizen	264
14.2.2	Additive Matrizen	266
14.3	Linkage-Algorithmen	268
14.4	Der Neighbour-Joining-Algorithmus	270
14.5	Parsimony-Methoden	272
14.6	Konstruktion eines Parsimony-Baumes	274
14.7	Maximum-Likelihood-Ansätze	275
14.7.1	Übergangswahrscheinlichkeiten für DNA-Sequenzen	275
14.7.2	Empirische Modelle der Protein-Evolution	276
14.7.3	Berechnen der Likelihood eines Baumes	278

14.7.4	Quartett-Puzzle	280
14.8	Grundannahmen phylogenetischer Algorithmen	283
14.9	Phylogenetische Analyse und statistische Bewertung	284
14.9.1	Verwenden von Outgroups	284
14.9.2	Das Bootstrap-Verfahren	284
14.10	Weitere phylogenetische Ansätze und Resultate	286
14.11	Zitierte Literatur	287
<b>15</b>	<b>Hidden-Markov-Modelle</b>	289
15.1	Eine Problem orientierte Einführung	290
15.2	Markov-Modelle	293
15.3	Ergodische Markovsche Ketten	300
15.3.1	Die Kolmogorov-Chapman-Gleichungen	300
15.3.2	Klassifikation der Zustände	301
15.3.3	Stationäre Verteilungen	306
15.3.4	Ergodizität von Quellen	308
15.3.5	Fazit	309
15.4	Niveau und Macht einfacher Tests	310
15.5	Exkurs: Grenzwertsätze	316
15.6	Diskrimination von CpG-Inseln	319
15.7	Ansätze zur Lokalisierung von CpG-Inseln	322
15.8	Der Begriff des Hidden-Markov-Modells	325
15.9	Wichtige Algorithmen für HMMs	328
15.9.1	Der Vorwärtsalgorithmus	329
15.9.2	Der Viterbi-Algorithmus	332
15.9.3	Der Rückwärtsalgorithmus	336
15.9.4	Die <i>A-posteriori</i> -Wahrscheinlichkeit der Zustände	337
15.10	Das zeitweise unehrliche Casino	339
15.11	Das Rekonstruktionsproblem für HMMs	342
15.11.1	Ein Maximum-Likelihood-Schätzer	342
15.11.2	Der Baum-Welch-Algorithmus zur Parameterschätzung	345
15.12	Zitierte Literatur	350
<b>16</b>	<b>Profil-HMMs zur Modellierung von Proteinfamilien</b>	351
16.1	Profil-HMMs	353
16.2	Viterbi-Pfade in Profil-HMMs	356
16.3	Eine Lösung des Anfrageproblems	361
16.4	Vorwärts- und Rückwärtsvariablen	362
16.5	Vom MSA zum Profil-HMM	366
16.6	Zitierte Literatur	369
<b>17</b>	<b>Bedingte Markovsche Zufallsfelder</b>	371
17.1	Markierungsprobleme und ME-Prinzip	372
17.1.1	Umfang eines Markierungsproblems	373
17.1.2	Merkmale	374

17.1.3	Maximierung der bedingten Entropie als Induktionsprinzip	375
17.1.4	ML-Parameterbestimmung	377
17.2	Der Satz von Hammersley und Clifford	378
17.3	IIS-Algorithmus	380
17.4	Linien-CRFs	383
17.4.1	Precomputing	383
17.4.2	Inferenz	385
17.4.3	Training: Umsetzung des IIS-Algorithmus	386
17.5	Zitierte Literatur	390
<b>18</b>	<b>Vorhersage der Sekundärstruktur</b>	391
18.1	Vorhersage der Proteinsekundärstruktur	392
18.1.1	Erste Ansätze: Chou-Fasman	392
18.1.2	PHD – Profil basierte Vorhersage	394
18.1.2.1	Vorgehensweise in PHD	394
18.1.2.2	Die Entwicklung und Validierung der Konformation von PHD	396
18.1.2.3	Trainieren der neuronalen Netze	397
18.1.2.4	Validierung mit <i>Leave-one-out</i> -Verfahren	397
18.2	Vorhersage der RNA-Sekundärstruktur	399
18.2.1	RNA-Sequenzen und -Strukturen	400
18.2.2	Freie Energie und Strukturen	401
18.2.3	Vorhersage der Sekundärstruktur durch Energieminimierung	403
18.2.4	Strukturen mit Schleifen	404
18.2.4.1	Berücksichtigung von Stacking-Interaktionen	405
18.2.4.2	Rekursionsgleichungen mit Stacking-Interaktionen	406
18.2.5	STAR: Vorhersage der Sekundärstruktur unter Verwendung eines genetischen Algorithmus	407
18.2.5.1	Erste Version des Modells	407
18.2.5.2	Zweite Version: Modellierung der RNA-Faltung	409
18.2.5.3	Ergebnisse	410
18.2.6	Weitere Verfahren zur Vorhersage von Strukturen mit Pseudoknoten	410
18.3	Zitierte Literatur	411
<b>19</b>	<b>Vergleich von Protein-3D-Strukturen</b>	413
19.1	Vergleich zweier Protein-3D-Strukturen	413
19.2	Superposition von Protein-3D-Strukturen	415
19.3	SAP: Vergleich von 3D-Strukturen mithilfe von Vektorbündeln	416
19.4	Simulated Annealing	419
19.5	Superposition mithilfe von DALI	422
19.5.1	Scores für Substrukturen	423
19.5.2	Alignieren von Substrukturen	424
19.6	TM-Align	425
19.7	Zitierte Literatur	427

<b>20</b>	<b>Homologiemodellierung und Vorhersage der Protein-3D-Struktur</b>	429
20.1	Verwenden von Threading-Verfahren	431
20.2	Eine Profil-Methode: 3D-1D-Profile	433
20.2.1	Bestimmen der Umgebungen	433
20.2.2	Generieren eines 3D-1D-Profiles	435
20.3	Wissensbasierte Kraftfelder	438
20.3.1	Theoretische Grundlagen	441
20.3.2	Ableiten der Potenziale	443
20.4	GenThreader	445
20.5	3D-PSSM	445
20.5.1	Generieren einer Profil-Bibliothek	446
20.5.2	Erstellen einer 3D-PSSM	447
20.5.3	Prozessieren der Query	449
20.5.4	Strukturvorhersage	450
20.5.5	Beitrag individueller Parameter	452
20.6	HHsearch	453
20.6.1	Grundlagen des Alignments von Hidden-Markov-Ketten	453
20.6.2	Paarweises Alignment von HMMs	457
20.6.3	Performanz von HHsearch	458
20.6.4	Strukturvorhersage mit HHsearch	460
20.7	ROSETTA/ROBETTA	460
20.7.1	Energieterme	461
20.7.2	<i>De novo</i> Strukturvorhersage mit ROSETTA	464
20.7.3	Alternativen zur Fragmentinsertion	465
20.7.4	Modellieren strukturell variabler Regionen in Homologiemodellen	466
20.8	Weitere Ansätze	467
20.9	Zitierte Literatur	468
<b>21</b>	<b>Analyse integraler Membranproteine</b>	471
21.1	Struktur integraler Membranproteine	472
21.2	Spezifische Probleme beim Sequenzvergleich	473
21.3	Vorhersage der Topologie von Helix-Bündeln	474
21.3.1	HMMTOP: das Topologiemodell	474
21.3.2	HMMTOP: Architektur des HMMs	476
21.4	Vorhersage der Topologie und Struktur von $\beta$ -Fässern	477
21.4.1	Architektur von TMBpro	478
21.4.2	Ausgabe und Performanz von TMBpro	479
21.5	Gegenwärtiger Stand bioinformatischer Methoden	480
21.6	Zitierte Literatur	480
<b>22</b>	<b>Entschlüsselung von Genomen</b>	481
22.1	Shotgun-Sequenzierung	484
22.2	Die Anzahl von Contigs beim Shotgun-Ansatz	485
22.3	Basecalling	486
22.4	Assemblieren von Teilesequenzen	488

22.4.1	Phase 1: Bestimmen überlappender Präfix-/Suffix-Regionen	489
22.4.2	Phase 2: Erzeugen von Contigs	490
22.4.3	Phase 3: Generieren der Konsensus-Sequenz	491
22.5	Annotation kompletter Genome	492
22.6	Metagenomik	496
22.6.1	Spezielle Anforderungen an die Bioinformatik	496
22.6.2	Minimalanforderungen für Metagenom-Annotation	497
22.7	Zitierte Literatur	498
<b>23</b>	<b>Auswertung von Genexpressionsdaten</b>	501
23.1	DNA-Chip-Technologie	501
23.1.1	Datenbanken für Genexpressionsdaten	503
23.1.2	Grenzen der Technologie	504
23.2	Bioinformatische Analyse von DNA-Chip-Signalen	505
23.2.1	Quantifizierung von Expressionswerten	505
23.2.2	Normalisierung und Datenreduktion	506
23.2.3	Normalisierung über Replikate	510
23.3	Identifizieren differentiell exprimierter Gene	511
23.4	Metriken zum Vergleich von Expressionsdaten	511
23.5	Algorithmen für die Analyse kompletter DNA-Chip-Datensätze	513
23.5.1	Anwendung von Clusterverfahren auf Genexpressionsdaten	514
23.5.2	Validierung und Alternativen	514
23.6	Hauptkomponentenanalyse	515
23.7	Biclusterverfahren	517
23.7.1	Ein Beispiel für Biclusterverfahren: ISA	518
23.7.2	Der Signatur-Algorithmus	519
23.7.3	Iterative Optimierung	522
23.8	Grenzen und Alternativen	524
23.9	Genexpressions-Profilierung	524
23.10	Wärmekarten	525
23.10.1	Der klassische Ansatz	526
23.10.2	Kombination von Datenquellen mithilfe von ClusCor	527
23.11	Informationsgewinnung für systembiologische Fragestellungen	528
23.11.1	Bündelung von Datenbankinformation	529
23.11.2	Statistische Analyse der Termverteilung	529
23.11.3	Verwendbarkeit des Verfahrens	530
23.12	Zitierte Literatur	530
<b>24</b>	<b>Analyse von Protein-Protein-Interaktionen</b>	533
24.1	Biologische Bedeutung des Interaktoms	534
24.2	Methoden zum Bestimmen des Interaktoms	534
24.3	Anforderungen an Datenbanksysteme	536
24.4	Analyse des Genominhaltes	537
24.4.1	Genfusion	538
24.4.2	Phyletische Muster	539

24.4.3	Analyse von Genfolgen	540
24.4.4	Performanz Sequenz basierter Methoden	541
24.5	Bewertung von Codon-Häufigkeiten	542
24.6	Suche nach korrelierten Mutationen	543
24.6.1	Generieren von sortierten MSA-Paaren	544
24.6.2	Identifizieren korrelierter Mutationen	544
24.7	Vergleich phylogenetischer Bäume	545
24.7.1	Die Mirror-tree-Methode	546
24.7.2	Korrektur des Hintergrundsignals	547
24.8	Vorhersage des Interaktoms der Hefe mithilfe eines Bayesschen Klassifikators	548
24.9	Zitierte Literatur	553
<b>25</b>	<b>Zum Schluss</b>	555
25.1	Zitierte Literatur	559
	<b>Stichwortverzeichnis</b>	561