

## Inhaltsverzeichnis

**Vorwort** *XV*

**Teil I Grundlagen – Biologie und Datenbanken** *1*

<b>1</b>	<b>Biologische Grundlagen</b>	<i>3</i>
1.1	DNA	<i>3</i>
1.2	Genetischer Code und Genomkomposition	<i>5</i>
1.3	Transkription	<i>9</i>
1.4	RNA	<i>10</i>
1.5	Proteine	<i>11</i>
1.6	Peptidbindung	<i>12</i>
1.7	Konformation von Aminosäureseitenketten	<i>13</i>
1.8	Ramachandran-Plot	<i>14</i>
1.9	Hierarchische Beschreibung von Proteinstrukturen	<i>16</i>
1.10	Sekundärstrukturelemente	<i>16</i>
1.11	$\alpha$ -Helix	<i>17</i>
1.12	$\beta$ -Faltblätter	<i>17</i>
1.13	Supersekundärstrukturelemente	<i>18</i>
1.14	Proteindomänen	<i>19</i>
1.15	Proteinfamilien	<i>20</i>
1.16	Enzyme	<i>23</i>
1.17	Proteinkomplexe	<i>24</i>
1.18	Fachbegriffe	<i>26</i>
	Literatur	<i>28</i>
<b>2</b>	<b>Sequenzen und ihre Funktion</b>	<i>31</i>
2.1	Definitionen und Operatoren	<i>32</i>
2.2	DNA-Sequenzen	<i>33</i>
2.3	Protein-Sequenzen	<i>33</i>
2.4	Vergleich der Sequenzkomposition	<i>35</i>
2.5	Ontologien	<i>38</i>
2.6	Semantische Ähnlichkeit von GO-Termen	<i>41</i>

- 2.6.1 Bewertung mittels informationstheoretischer Ansätze 42
- 2.6.2 Vergleich mit einer graphentheoretischen Methode 43
- Literatur 46

### **3 Datenbanken 47**

- 3.1 Nukleotidsequenz-Datenbanken 48
- 3.2 RNA-Sequenz-Datenbanken 49
- 3.3 Proteinsequenz-Datenbanken 49
- 3.4 3D-Struktur-Datenbanken 50
- 3.5 SMART: Analyse der Domänenarchitektur 51
- 3.6 STRING: Proteine und ihre Interaktionen 52
- 3.7 SCOP: Strukturelle Klassifikation von Proteinen 53
- 3.8 Pfam: Kompilation von Proteinfamilien 55
- 3.9 COG und eggNOG: Gruppen orthologer Gene 56
- 3.10 Weitere Datenbanken 57
- Literatur 60

## **Teil II Lernen, Optimieren und Entscheiden 63**

### **4 Grundbegriffe der Stochastik 65**

- 4.1 Grundbegriffe der beschreibenden Statistik 66
- 4.2 Zufallsvariable, Wahrscheinlichkeitsmaß 68
- 4.3 Urnenexperimente und diskrete Verteilungen 70
- 4.4 Die Kolmogoroffschen Axiome 71
- 4.5 Bedingte Wahrscheinlichkeit, Unabhängigkeit, Satz von Bayes 73
- 4.6 Markov-Ketten 74
- 4.7 Erwartungswert, Varianz 74
- 4.8 Wichtige Wahrscheinlichkeitsverteilungen 75
- 4.8.1 Diskrete Verteilungen 75
- 4.8.2 Totalstetige Verteilungen 76
- 4.9 Schätzer 79
- 4.10 Grundlagen statistischer Tests 81
- 4.11 Eine optimale Entscheidungstheorie: Die Neyman-Pearson-Methode 82
- Literatur 84

### **5 Bayessche Entscheidungstheorie und Klassifikatoren 85**

- 5.1 Bayessche Entscheidungstheorie 85
- 5.1.1 Ein Beispiel: Klassifikation der Proteinoberfläche 86
- 5.1.2 Übergang zu bedingten Wahrscheinlichkeiten 87
- 5.1.3 Erweitern auf  $m$  Eigenschaften 89
- 5.2 Marginalisieren 91
- 5.3 Boosting 91
- 5.4 ROC-Kurven 94

5.4.1	Bewerten von Fehlklassifikationen	94
5.4.2	Aufnehmen einer ROC-Kurve	94
5.5	Testmethoden für kleine Trainingsmengen	97
	Literatur	99
<b>6</b>	<b>Klassische Cluster- und Klassifikationsverfahren</b>	<b>101</b>
6.1	Metriken und Clusteranalyse	102
6.2	Das mittlere Fehlerquadrat als Gütemaß	102
6.3	Ein einfaches iteratives Clusterverfahren	104
6.4	$k$ -Means-Clusterverfahren	105
6.5	Hierarchische Clusterverfahren	108
6.6	Nächster-Nachbar-Klassifikation	109
6.7	$k$ nächste Nachbarn	110
	Literatur	111
<b>7</b>	<b>Neuronale Netze</b>	<b>113</b>
7.1	Architektur von neuronalen Netzen	113
7.2	Das Perzeptron	114
7.3	Modellieren Boolescher Funktionen	116
7.4	Lösbarkeit von Klassifikationsaufgaben	116
7.5	Universelle Approximation	119
7.6	Lernen in neuronalen Netzen	121
7.7	Der Backpropagation-Algorithmus	122
7.8	Codieren der Eingabe	125
7.9	Selbstorganisierende Karten	126
	Literatur	128
<b>8</b>	<b>Genetische Algorithmen</b>	<b>131</b>
8.1	Objekte und Funktionen	133
8.2	Beschreibung des Verfahrens	135
8.3	Der Begriff des Schemas	136
8.4	Dynamik der Anzahl von Schemata	137
8.5	Codieren der Problemstellung	139
8.6	Genetisches Programmieren	139
	Literatur	141
<b>Teil III Algorithmen und Modelle der Bioinformatik</b>		<b>143</b>
<b>9</b>	<b>Paarweiser Sequenzvergleich</b>	<b>145</b>
9.1	Dotplots	147
9.1.1	Definition	147
9.1.2	Beispiel	148
9.1.3	Implementierung	149
9.1.4	Abschätzen der Laufzeit	149

- 9.1.5 Anwendungen 150
- 9.1.6 Einschränkungen und Ausblick 152
- 9.2 Entwickeln eines optimalen Alignmentverfahrens 154
- 9.2.1 Paarweise und multiple Sequenzalignments 156
- 9.2.2 Dynamisches Programmieren 156
- 9.2.3 Distanzen und Metriken 158
- 9.2.4 Die Minkowski-Metrik 159
- 9.2.5 Die Hamming-Distanz 159
- 9.3 Levenshtein-Distanz 161
- 9.3.1 Berechnungsverfahren 163
- 9.3.2 Ableiten des Alignments 165
- 9.4 Bestimmen der Ähnlichkeit von Sequenzen 165
- 9.4.1 Globales Alignment 167
- 9.4.2 Lokales Sequenzalignment 167
- 9.5 Optimales Bewerten von Lücken 168
- 9.5.1 Eigenschaften affiner Kostenfunktionen 169
- 9.5.2 Integration in Algorithmen 170
- 9.6 Namensgebung 171
- Literatur 172
  
- 10 Sequenzmotive 173**
- 10.1 Signaturen 174
- 10.2 Die PROSITE-Datenbank 175
- 10.3 Die BLOCKS-Datenbank 175
- 10.4 Sequenzprofile 176
- 10.5 Scores für Promotorsequenzen 178
- 10.6 Möglichkeiten und Grenzen profilbasierter Klassifikation 178
- 10.7 Sequenz-Logos 179
- 10.8 Konsensus-Sequenzen 180
- 10.9 Sequenzen niedriger Komplexität 181
- 10.10 Der SEG-Algorithmus 182
- Literatur 184
  
- 11 Scoring-Schemata 187**
- 11.1 Theorie von Scoring-Matrizen 188
- 11.2 Algorithmenbedingte Anforderung 190
- 11.3 Identitätsmatrizen 191
- 11.4 PAM-Einheit 191
- 11.5 PAM-Matrizen 192
- 11.6 Ein moderner PAM-Ersatz: Die JTT-Matrix 193
- 11.7 BLOSUM-Matrizen 195
- 11.8 Matrix-Entropie 198
- 11.9 Scoring-Schemata und Anwendungen 199
- 11.10 Flexible Erweiterung: Scoring-Funktionen 200
- Literatur 201

- 12 FASTA und die BLAST-Suite 203**
  - 12.1 FASTA 204
    - 12.1.1 Programmablauf 204
    - 12.1.2 Statistische Bewertung der Treffer 206
  - 12.2 BLAST 209
    - 12.2.1 Konzepte und Umsetzung 210
    - 12.2.2 Statistik von Alignments 212
    - 12.2.3 Ausgabe der Treffer 216
  - 12.3 Vergleich der Empfindlichkeit von FASTA und BLAST 217
  - 12.4 Ansätze zur Performanzsteigerung 218
  - 12.5 Profilbasierter Sequenzvergleich 219
  - 12.6 PSI-BLAST 219
  - 12.7 Sensitivität verschiedener Sequenzvergleichsmethoden 222
  - 12.8 Vergleich von Profilen und Konsensus-Sequenzen 224
  - 12.9 DELTA-BLAST 225
  - Literatur 228
  
- 13 Multiple Sequenzalignments und Anwendungen 229**
  - 13.1 Berechnen von Scores für multiple Sequenzalignments 231
  - 13.2 Iteratives Berechnen eines Alignments 231
  - 13.3 ClustalW: Ein klassischer Algorithmus 233
    - 13.3.1 Grundlegende Konzepte 233
    - 13.3.2 Algorithmus 233
    - 13.3.3 Ein Beispiel: MSA für Trypsin-Inhibitoren 234
  - 13.4 T-Coffee 236
  - 13.5 M-Coffee und 3D-Coffee 239
  - 13.6 Alternative Ansätze 241
  - 13.7 Alignieren großer Datensätze 241
  - 13.8 Charakterisierung von Residuen mithilfe von Alignments 242
    - 13.8.1 Entwickeln der Scoring-Funktion 244
    - 13.8.2 *FRpred*: Vorhersage funktionell wichtiger Residuen 245
    - 13.8.3 *SDPpred*: Vergleich homologer Proteine mit unterschiedlicher Spezifität 246
  - 13.9 Alignment von DNA- und RNA-Sequenzen 247
  - Literatur 248
  
- 14 Grundlagen phylogenetischer Analysen 251**
  - 14.1 Einteilung phylogenetischer Ansätze 255
  - 14.2 Distanzbasierte Verfahren 256
    - 14.2.1 Ultrametrische Matrizen 256
    - 14.2.2 Additive Matrizen 258
  - 14.3 Linkage-Algorithmen 259
  - 14.4 Der Neighbour-Joining-Algorithmus 261
  - 14.5 Parsimony-Methoden 263
  - 14.6 Maximum-Likelihood-Ansätze 266

14.6.1	Übergangswahrscheinlichkeiten für DNA-Sequenzen	266
14.6.2	Empirische Modelle der Protein-Evolution	267
14.6.3	Berechnen der Likelihood eines Baumes	268
14.6.4	Quartett-Puzzle: Heuristik zum Finden einer Topologie	271
14.7	Grundannahmen phylogenetischer Algorithmen	274
14.8	Statistische Bewertung phylogenetischer Bäume	275
14.8.1	Verwenden von Outgroups	275
14.8.2	Bootstrap-Verfahren und posterior Wahrscheinlichkeiten	276
14.9	Alternativen und Ergebnisse	277
	Literatur	278
<b>15</b>	<b>Markov-Ketten und Hidden-Markov-Modelle</b>	<b>281</b>
15.1	Ein epigenetisches Signal: CpG-Inseln	281
15.2	Finite Markov-Ketten	282
15.3	Kombination zweier Ketten zu einem Klassifikator	283
15.4	Genvorhersage mithilfe inhomogener Ketten	286
15.5	Hidden-Markov-Modelle	288
15.6	Der Viterbi-Pfad	292
15.7	Ein HMM zur Erkennung von CpG-Inseln	294
15.8	Der Vorwärts- und der Rückwärts-Algorithmus	294
15.9	Schätzen von Parametern	297
15.10	Der Baum-Welch-Algorithmus	298
15.11	Entwurf von HMMs	299
15.12	Verwendung und Grenzen von HMMs	301
15.13	Wichtige Eigenschaften von Markov-Ketten	302
15.14	Markov-Ketten-Monte-Carlo-Verfahren	304
15.14.1	Monte-Carlo-Integration	305
15.14.2	Metropolis-Hastings-Algorithmus	305
15.14.3	Simulated Annealing	307
15.14.4	Gibbs-Sampler	307
15.15	Weitere Anwendungen von Markov-Ketten	308
	Literatur	310
<b>16</b>	<b>Profil-HMMs</b>	<b>313</b>
16.1	HMM-Struktur zur Beschreibung von Proteinfamilien	314
16.2	Suche nach homologen Sequenzen	317
16.3	Modellbau mit Profil-HMMs	320
16.4	Approximieren von Wahrscheinlichkeitsdichten	324
16.5	HHsearch: Vergleich zweier Profil-HMMs	330
16.5.1	Grundlagen des Alignments von zwei Hidden-Markov-Ketten	331
16.5.2	Paarweises Alignment von HMMs	334
16.5.3	Performanz von HHsearch	336
16.5.4	Strukturvorhersage mit HHsearch	337
	Literatur	338

- 17 Support-Vektor-Maschinen 339**
  - 17.1 Beschreibung des Klassifikationsproblems 340
  - 17.2 Lineare Klassifikatoren 341
  - 17.3 Klassifizieren mit großer Margin 345
  - 17.4 Kernel-Funktionen und Merkmalsräume 347
  - 17.5 Implizite Abbildung in den Merkmalsraum 348
  - 17.6 Eigenschaften von Kernel-Funktionen 350
  - 17.7 Häufig verwendete Kernel-Funktionen 351
  - 17.8 Aus Merkmalen abgeleitete Kernel-Funktionen 353
  - 17.9 Support-Vektor-Maschinen in der Anwendung 356
  - 17.10 Multiklassen SVMs 359
  - 17.11 Theoretischer Hintergrund 360
    - Literatur 363
  
- 18 Vorhersage der Sekundärstruktur 365**
  - 18.1 Vorhersage der Proteinsekundärstruktur 366
    - 18.1.1 Ein früher Ansatz: Chou-Fasman-Verfahren 367
    - 18.1.2 PHD: Profilbasierte Vorhersage 367
  - 18.2 Vorhersage der RNA-Sekundärstruktur 373
    - 18.2.1 RNA-Sequenzen und -Strukturen 374
    - 18.2.2 Freie Energie und Strukturen 375
    - 18.2.3 Sekundärstrukturvorhersage durch Energieminimierung 377
    - 18.2.4 Strukturen mit Schleifen 378
    - 18.2.5 STAR: Einbinden eines genetischen Algorithmus 380
    - 18.2.6 MEA-Verfahren zur Vorhersage von Strukturen mit Pseudoknoten 383
    - 18.2.7 Strukturvorhersage mithilfe von multiplen Sequenzalignments 386
      - Literatur 388
  
- 19 Vergleich von Protein-3D-Strukturen 389**
  - 19.1 Grundlagen des Strukturvergleichs 390
  - 19.2 Superposition von Protein-3D-Strukturen 392
  - 19.3 SAP: Vergleich von 3D-Strukturen mit Vektorbündeln 393
  - 19.4 Simulated Annealing 395
  - 19.5 Superposition mithilfe von DALI 398
    - 19.5.1 Scores für Substrukturen 399
    - 19.5.2 Alignieren von Substrukturen 400
  - 19.6 TM-Align 400
  - 19.7 DeepAlign 402
  - 19.8 Multiple Superpositionen 408
    - Literatur 409
  
- 20 Vorhersage der Protein-3D-Struktur 411**
  - 20.1 Threading-Verfahren 416
  - 20.2 3D-1D-Profil: Profilbasiertes Threading 418

20.2.1	Bestimmen der lokalen Umgebung	418
20.2.2	Erzeugen eines 3D-1D-Profiles	420
20.3	Wissensbasierte Kraftfelder	423
20.3.1	Theoretische Grundlagen	424
20.3.2	Ableiten der Potenziale	427
20.4	Rotamerbibliotheken	428
20.5	MODELLER	432
20.6	ROSETTA/ROBETTA	436
20.6.1	Energieterme und ihre Verwendung	437
20.6.2	<i>De-novo</i> -Strukturvorhersage mit ROSETTA	438
20.6.3	Verfeinerung der Fragmentinsertion	440
20.6.4	Modellieren strukturell variabler Regionen	441
20.7	Alternative Modellieransätze	443
20.8	Verify-3D: Bewerten der Modellqualität	444
	Literatur	445
<b>21</b>	<b>Analyse integraler Membranproteine</b>	<b>447</b>
21.1	Architektur integraler Membranproteine	448
21.2	Spezifische Probleme beim Sequenzvergleich	450
21.3	Vorhersage der Topologie von Helix-Bündeln	450
21.3.1	HMMTOP	450
21.3.2	MEMSAT-SVM	453
21.3.3	Ein Meta-Server: TOPCONS	454
21.4	Vorhersage der Struktur von $\beta$ -Fässern	454
21.4.1	TMBpro	454
21.4.2	BOCTOPUS	456
21.5	Alternative Ansätze und Homologiemodellierung	457
21.6	Gegenwärtiger Stand bioinformatischer Methoden	458
	Literatur	459
<b>22</b>	<b>Entschlüsselung von Genomen</b>	<b>461</b>
22.1	Shotgun-Sequenzierung	464
22.2	Erwartete Anzahl von Contigs beim Shotgun-Ansatz	465
22.3	Basecalling und Sequenzqualität	467
22.4	Assemblieren von Teilsequenzen: Klassischer Ansatz	468
22.4.1	Phase eins: Bestimmen überlappender Präfix/Suffix-Regionen	469
22.4.2	Phase zwei: Erzeugen von Contigs	471
22.4.3	Phase drei: Generieren der Konsensus-Sequenz	471
22.5	Neue Herausforderung: Assemblieren kurzer Fragmente	473
22.6	Annotation kompletter Genome	476
22.7	Metagenomik	481
22.7.1	Spezielle Anforderungen an die Bioinformatik	482
22.7.2	Minimalanforderungen für die Metagenom-Annotation	484
	Literatur	484

- 23 Auswertung von Genexpressionsdaten 487**
- 23.1 DNA-Chip-Technologie 487
- 23.1.1 Datenbanken für Genexpressionsdaten 489
- 23.1.2 Grenzen der Technologie 490
- 23.2 Analyse von DNA-Chip-Signalen 490
- 23.2.1 Quantifizierung von Expressionswerten 491
- 23.2.2 Normalisieren und Datenreduktion 492
- 23.2.3 Normalisieren über Replikate 495
- 23.3 Identifizieren differenziell exprimierter Gene 496
- 23.4 Metriken zum Vergleich von Expressionsdaten 497
- 23.5 Analyse kompletter DNA-Chip-Datensätze 498
- 23.5.1 Anwendung von Clusterverfahren 498
- 23.5.2 Validierung und Alternativen 499
- 23.6 Hauptkomponentenanalyse 500
- 23.7 Biclusterverfahren 502
- 23.7.1 ISA: Ein performantes Biclusterverfahren 502
- 23.7.2 Der Signatur-Algorithmus 503
- 23.7.3 Iterative Optimierung 506
- 23.7.4 QUBIC: Ein graphenbasiertes Biclusterverfahren 508
- 23.8 Grenzen und Alternativen bei der Expressionsanalyse 509
- 23.9 Genexpressions-Profiling 509
- 23.10 Visualisieren mithilfe von Wärmekarten 510
- 23.10.1 Der klassische Ansatz 510
- 23.10.2 ClusCor: Kombination verschiedener Datenquellen 511
- 23.11 Datenaufbereitung für systembiologische Fragestellungen 512
- 23.11.1 Bündelung von Datenbankinformation 513
- 23.11.2 Statistische Analyse der Termverteilung 515
- 23.11.3 Verwendbarkeit der Verfahren 515
- Literatur 516
  
- 24 Analyse von Protein-Protein-Interaktionen 519**
- 24.1 Biologische Bedeutung des Interaktoms 519
- 24.2 Methoden zum Bestimmen des Interaktoms 520
- 24.3 Analyse des Genominhaltes 521
- 24.3.1 Genfusion 522
- 24.3.2 Phyletische Muster 523
- 24.3.3 Analyse von Genfolgen 524
- 24.3.4 Performanz sequenzbasierter Methoden 525
- 24.4 Bewerten von Codonhäufigkeiten 526
- 24.5 Suche nach korrelierten Mutationen 527
- 24.5.1 Erzeugen sortierter MSA-Paare 527
- 24.5.2 Identifizieren korrelierter Mutationen 528
- 24.6 Vergleich phylogenetischer Bäume 529
- 24.6.1 Die mirror-tree-Methode 529
- 24.6.2 Korrektur des Hintergrundsignals 531

24.7	Vorhersage des Interaktoms der Hefe	532
24.8	Protein-Protein-Interaktionsvorhersagen	535
24.8.1	Vorhersagen basierend auf Strukturinformation	536
24.8.2	PrePPI: Integration zusätzlicher Merkmale	538
	Literatur	542
<b>25</b>	<b>Big Data: Herausforderungen und neue Möglichkeiten</b>	<b>545</b>
25.1	Klassifikation mit Random Forests	547
25.1.1	Entscheidungsbäume	547
25.1.2	Berechnen der Topologie	549
25.1.3	RF-Algorithmus	551
25.1.4	Theoretische Klassifikationsleistung eines RFs	553
25.1.5	Problemlösungen für konkrete Anwendungen	554
25.1.6	Auswahl informativer Eigenschaften	555
25.1.7	Bioinformatische Anwendungen	557
25.2	Sequenzbasierte Vorhersage der Protein-3D-Struktur	558
25.2.1	Experimentelle Proteinstrukturaufklärung	559
25.2.2	Berechnen von Kovariationssignalen	560
25.2.3	PSICOV: Vorhersage räumlich benachbarter Residuen-Paare	563
25.2.4	Vorhersage der 3D-Struktur mithilfe von Kontaktinformation	565
25.2.5	Alternative Nutzung von Kopplungssignalen	565
25.3	Berechnen einer Feinstruktur großer Proteinfamilien	566
25.3.1	MCL: Clustern mithilfe stochastischer Matrizen	567
25.3.2	Cytoscape: Visualisierung von Netzwerk-Clustern	569
25.4	Positionierung von Nukleosomen	570
25.4.1	Chromatin und Nukleosomen	571
25.4.2	NucleoFinder: Statistischer Ansatz zur Vorhersage von Nukleosomen-Positionen	572
25.5	Analyse des menschlichen Genoms mithilfe von ENCODE-Daten	576
25.5.1	Datentypen	577
25.5.2	Genom-Browser	579
	Literatur	581
<b>26</b>	<b>Zum Schluss</b>	<b>585</b>
26.1	Informatik in schwierigem Umfeld	585
26.2	Ungelöste Probleme und Herausforderungen	587
	Literatur	589
	<b>Index</b>	<b>591</b>