# Problemstellungen der Bioinformatik

Proseminar im Grundstudium, Sommersemester 2008

## Themen

# 1 Sequenzierung

Inhalt: Erklären der üblichen Verfahren zur Sequenzierung von DNA und RNA (chemische Methode, Kettenabbruchmethode) und der Aminosäuresequenzierung. Beschreiben der Funktionsweise der Gelelektrophorese. Weitere eventuelle Themen: 2D-Gelelektrophorese und PAGE.

Literatur: [Alp98]

# 2 Genetische und physikalische Karten

Inhalt:

- Genetische Karten: Bestimmung der (relativen) Lage von Genen auf den Chromosomen (nur kurz)
- Physikalische Karten: Bestimmen der Lage von größeren DNA-Teilen und/oder Markern

  Verfahren bei der Annahme fehlerfreier Daten und unter Berücksichtigung von Fehlern (vor allem: double digest, single digest, restriction site mapping, hybridisation mapping)

Literatur: [Cas]; [Wat95]: Kap. 6; [Set97]: Kap. 1.5 und 5; [Gus97]: Teil aus Kap. 16, [Böc03]: Kap. 7

Weitere Links: `http://www.cs.technion.ac.il/Labs/cbl/teaching/bab/`

# 3 Fragmentassemblierung: die shot gun Methode

Inhalt: Zusammensetzen von Teilsequenzen eines größeren DNA-Stücks unter Berücksichtigung von Fehlern sowie unter Annahme der Korrektheit.

(heuristischer und exakter Algorithmus)

Literatur: [Set97]: Kap. 4; [Cas]; [Böc03], Kap. 8

# 4 Human Genome Project: Sequenzierung des menschlichen Genoms

Als Startpunkt: [Int01], [Mur02], [Ven00]

# 5 ENCODE Projekt

Inhalt: Beschreibung des Projektes: Resequenzierung eines Teils des Humangenoms und funktionelle Annotation. Ziel: Gesamtgenom und -transkriptom charakterisieren.

Literatur: [Con04] [Pen07]

# 6 Phylogenetische Bäume

Inhalt Ermitteln von Stammbäumen, die dabei auftretenden Problem und approximative Lösungen dafür.

Literatur [Set97]: Kap. 6; [Gus97]; [Wat95]

# 7 Einführung Massenspektrometrie

- GC/LC-MS Technik, [Leh96]: Kap. 1 und 4
- Chromatigraphische Trennung
- Ionisierungsmethoden
- Detektoren
- AMDIS Software `http://www.amdis.net/`, [Ste99],[Dav04]

# 8 Signalverarbeitung und Alignment von LC-MS-Daten

- XCMS [Smiss]

# 9 Metaboliten Identifikation mit hochauflösenden Massenspektrometern

- MSMS Technik, Fragmentmuster, Substrukturen [Leh96]: Kap. 1.3
- Arabidopsis Profiling, MS/MS zur Identifizierung [RL04]
- Exakte Masse, Isotopenmuster

# 10 Proteindatenbanken

Literatur PDB: [Ber77]; [Sus98]
SWISS-PROT : [A.04], `http://www.expasy.org/sprot/`

# 11 Vorhersage von Proteinstrukturen

Inhalt Allgemein das Problem der Vorhersage einer Struktur aus einer bekannten Sequenz und aktuelle Lösungsansätze (homoloqy modeling, fold recognition, ab initio)

Literatur *Biochemische Hintergünde:*
[Bro69]
*Allgemein zur Vorhersageproblematik:*
[Kön97]; [Len96]
[Clo00]: Teil aus Kap. 6, [Gla95]: Kap. 9.III

Weitere Links `www.tcm.phy.cam.ac.uk/~mmlk2/report13/report13.html`

# 12 Protein-Ligand-Docking

Literatur FlexX `http://cartan.gmd.de/flexx/`; [Jon97]

# 13 Microarrays: Datengewinnung, Vorverarbeitung und Normalisierung

Inhalt Vorstellung verschiedener Arten von Microarrays, ihrer Herstellung und der Datengewinnung.

weitere Themen zur Vertiefung:

- Erläuterung eines genetischen Algorithmus' zur Bestimmung von Sonden
- Normalisierung der Daten, sodass man mehrere Microarrays miteinander vergleichen kann

Literatur [Bow99]; [Kel98]; [Hac99]; Das Affymetrix Benutzerhandbuch

Weitere Links `www.affymetrix.com`

# 14 Analyse von Microarrays: Clustern

Inhalt Distanzmaße und Linkage-Verfahren, sowie hierarchisches Clustern am Beispiel des Eisen-Programms. Weitere Themen: SOMs oder k-means.

Literatur Anwendungen: [Eis98]; [Tam99]

# 15 Tiling Arrays

Inhalt ChIP-on-chip, Tiling arrays zur Transkriptomüberwachung, DNAse-Chips um offene Chromatinstrukturen zu erkennen. Biologische/biochemische Methoden, Zielstellungen und Ansatzpunkte für die Bioinformatik.

Literatur Als Startpunkt: [Moc05] [Liu07]

# 16 Klassifikatoren: Support Vector Machines

Inhalt Vorstellung der Konzepte von Support Vector Machines, Klassifikationsregel, Kernels. Anwendung auf biologische Daten (z.B. Expressionsdaten, (DNA-) Sequenzdaten).

Literatur [Bur98, Bro99, Mei04]

# 17 Sequenz-Alignment

Inhalt Vorstellung von Algorithmen (dynamisches Programmieren) zur Berechnung von lokalen und globalen Alignments zwischen zwei Sequenzen und mögliche Bewertungsfunktionen (Distanzen, Ähnlichkeiten). Weitere Themen: Multiple Alignments (zwischen mehr als zwei Sequenzen) und verwendete Heuristiken.

Literatur [Set97]: Kap. 3; [Gus97]; [Wat95]

# 18 FASTA und BLAST

Inhalt Vorstellung von Sequenzdatenbanken und Alignments mit FASTA und BLAST und Erläuterung der dort verwendeten Heuristiken und Bewertungsmatrizen/-verfahren PAM, BLOSUM.

Literatur [Set97]: Kap. 3.5; [Gus97]

# 19 Statistische Modellierung von Sequenzdaten

Inhalt Statistische Modellierung von Sequenzdaten mit *position weight matrices* (PWMs) und *weight array models* (WAMs). Darstellung von Konsensussequenzen. Weitere Themen: Klassifikation mit statistischen Modellen.

Literatur [Sal97, Zha93, Sta84]

Weitere Links `http://www.gene-regulation.com/cgi-bin/pub/programs/match/bin/match.cgi`

# 20 Phylogenetic Footprinting

Inhalt Problemstellungen für Phylogenetic Footprinting, konservierte vs. nichtkonservierte DNA, Anwendung zum Finden von Transkriptionsfaktorbindungsstellen, Fehlerquellen

Literatur [Zha03]; [Bof03]; [Bla02]

# 21 DNA-Computing

[Bra02, Rub00]

Bei der angegebenen Literatur handelt es sich um eine "Basisausrüstung" – es können und sollen auch andere Quellen hinzugezogen werden.

# Literatur


[A.04] B. A., B. B., F. S., G. E.: *Swiss-Prot: Juggling between evolution and stability*, *Briefings in Bioinformatics*, Bd. 5, 2004, S. 39–55.

[Alp98] L. Alphey: *DNA-Sequenzierung*, Spektrum Akademischer Verlag, Heidelberg, Berlin, 1998.

[Ber77] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. J. Meyer, M. Brice, J. Rodgers, O. Kennard, T. Shimanouchi, M. Tasumi: *The Protein Data Bank: a computer–based archival file for macromolecular structures*, *Journal of Molecular Biology*, Bd. 112, 1977, S. 535–542.

[Bla02] M. Blanchette, M. Tompa: *Discovery of Regulatory Elements by a Computational Method for Phylogenetic Footprinting*, *Genome Res.*, Bd. 12, Nr. 5, 2002, S. 739–748.

[Böc03] H.-J. Böckenhauer, D. Bongartz: *Algorithmische Grundlagen der Bioinformatik*, Teubner, 2003.

[Bof03] D. Boffelli, J. McAuliffe, D. Ovcharenko, K. D. Lewis, I. Ovcharenko, L. Pachter, E. M. Rubin: *Phylogenetic Shadowing of Primate Sequences to Find Functional Regions of the Human Genome*, *Science*, Bd. 299, Nr. 5611, 2003, S. 1391–1394.

[Bow99] D. Bowtell: *Options available - from start to finish - for obtaining expression data by microarray*, *Nature Genetics Supplement*, Bd. 21, 1999, S. 25–32.

[Bra02] R. S. Braich, N. Chelyapov, C. Johnson, P. W. Rothemund, L. Adleman: *Solution of a 20-Variable 3-SAT Problem on a DNA Computer*, *Science*, Bd. 296, 2002, S. 499–502.

[Bro69] W. J. Browne, A. C. T. North, D. C. Phillips: *A Possible Three-dimensional Structure of Bovine α-Lactalbumin based on that of Hen's Egg-White Lysozyme*, *Journal of Molecular Biology*, Bd. 42, 1969, S. 65–86, Historisches Paper mit handgemachter Homologievorhersage, Drahtmodell und Stereophotos.

[Bro99] M. Brown, W. Grundy, D. Lin, N. Christianini, C. Sugnet, M. Jr, D. Haussler: *Support vector machine classification of microarray gene expression data*, 1999.

[Bur98] C. J. C. Burges: *A Tutorial on Support Vector Machines for Pattern Recognition*, *Data Mining and Knowledge Discovery*, Bd. 2, Nr. 2, 1998, S. 121–167.

[Cas] D. Casey: *Primer on Molecular Gentics*, http://www.ornl.gov/sci/techresources/Human_Genome/pul

[Clo00] P. Clote, R. Backofen: *Computational Molecular Biology*, Wiley, 2000.

[Con04] T. E. P. Consortium: *The ENCODE (ENCyclopedia Of DNA Elements) Project*, *Science*, Bd. 306, Nr. 5696, 2004, S. 636–640.

[Dav04] A. N. Davies: *The new Automated Mass Spectrometry Deconvolution and Identification System (AMDIS)*, *Spectroscopy Europe*, Bd. 10, Nr. 3, 2004, S. 22–26.

[Eis98] M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein: *Cluster analysis and display of genome-wide expression patterns*, *Proc. Natl. Acad. Sci. USA*, Bd. 95, 1998, S. 14863–14868.

[Gla95] J. A. Glasel (Hrsg.): *Introduction to biophysical methods for protein and nucleic acid research*, Academic Press, 1995, Physikalische Beschreibung von Rntgenstrukturanalyse und Beschreibung von Faltungsvorhersage.

[Gus97] D. Gusfield: *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*, Cambridge University Press, Cambridge, 1997.

[Hac99] J. Hacia: *Resequencing and mutational analysis using oligonucleotide microarrays*, *Nature Genetics*, Bd. 21, 1999, S. 42–47.

[Int01] International Human Genome Sequencing Consortium: *Initial sequencing and analysis of the human genome*, *Nature*, Bd. 409, Nr. 6822, February 2001, S. 860–921.

[Jon97] G. Jones, P. Willett, R. C. Glen, A. Leach, R. Taylor: *Development and Validation of a Genetic Algorithm for Flexible Docking*, *Journal Molecular Biology*, Bd. 267, 1997, S. 727–748.

[Kel98] A. Kel, A. Ptitsyn, V. Babenko, S. Meier-Ewert, H. Lehrach: *A genetic algorithm for designing gene family-specific oligonucleotide sets used for hybridization: the G protein-coupled receptor protein superfamily*, *Bioinformatics*, Bd. 14, Nr. 3, 1998, S. 259–270.

[Kön97] R. König, T. Dandekar: *Computational methods for the prediction of protein folds*, *Biochemica et Biophysika acta*, Bd. 1343, Nr. 1, 1997, S. 1.

[Leh96] W. D. Lehmann: *Massenspektrometrie in der Biochemie*, Spektrum, 1996.

[Len96] T. Lengauer, R. Thiele, R. Zimmer: *Modellierung von Proteinstrukturen*, *Der GMD-Spiegel*, Bd. 2/3, 1996, S. 14–18.

[Liu07] X. S. Liu: *Getting Started in Tiling Microarray Analysis*, *PLoS Comput Biol*, Bd. 3, Nr. 10, Okt 2007, S. e183.

[Mei04] P. Meinicke, M. Tech, B. Morgenstern, R. Merkl: *Oligo kernels for datamining on biological sequences: a case study on prokaryotic translation initiation sites*, *BMC Bioinformatics*, Bd. 5, Nr. 1, 2004, S. 169.

[Moc05] T. C. Mockler, J. R. Ecker: *Applications of DNA tiling arrays for whole-genome analysis*, *Genomics*, Bd. 85, Nr. 1, 2005, S. 1–15.

[Mur02] R. Mural: *A Comparison of Whole-Genome Shotgun-Derived Mouse Chromosome 16 and the Human Genome*, *Science*, Bd. 296, Nr. 5573, May 2002, S. 1661 – 1671.

[Pen07] E. Pennisi: *GENOMICS: DNA Study Forces Rethink of What It Means to Be a Gene*, *Science*, Bd. 316, Nr. 5831, 2007, S. 1556–1557.

[RL04] E. v. Roepenack-Lahaye, T. Degenkolb, M. Zerjeski, M. Franz, U. Roth, L. Wessjohann, J. Schmidt, D. Scheel, S. Clemens: *Profiling of Arabidopsis Secondary Metabolites by Capillary Liquid Chromatography Coupled to Electrospray Ionization Quadrupole Time-of-Flight Mass Spectrometry*, *Plant Physiology*, Bd. 134, February 2004, S. 548–559.

[Rub00] A. J. Ruben, L. F. Landweber: *The past, present and future of molecular computing*, *Nature Reviews Molecular Cell Biology*, Bd. 1, 2000, S. 69–72.

[Sal97] S. Salzberg: *A method for identifying splice sites and translational start sites in eukaryotic mRNA*, *Computer Applications in Biosciences*, Bd. 13, Nr. 4, 1997, S. 365–376.

[Set97] J. Setubal, J. Meidanis: *Introduction to Computational Molecular Biology*, PWS Publishing, Boston, Mass., 1997.

[Smiss] C. Smith, E. Want, G. O'Maille, R. Abagyan, G. Siuzdak: *XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching and identification*, *Analytical Chemistry*, 2006 (in Press).

[Sta84] R. Staden: *Measurements of the effects that coding for a protein has on a DNA sequence and their use for finding genes*, *Nucleic Acids Research*, Bd. 12, 1984, S. 789–800.

[Ste99] S. E. Stein: *An Integrated Method for Spectrum Extraction and Compound Identification from GC/MS Data*, *Journal of the American Society of Mass Spectrometry*, Bd. 10, 1999, S. 770–781.

[Sus98] J. L. Sussman, D. Lin, J. Jiang: *Protein Data Bank (PDB): Database of Three-Dimensional Structural Information of Biological Macromolecules*, *Acta Crystallographica Section D Biological Crystallography*, Bd. 6, 1998, S. 1078–.

[Tam99] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, T. R. Golub: *Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation*, *Proc. Natl. Acad. Sci. USA*, Bd. 96, Nr. 6, 1999, S. 2907–2912.

[Ven00] C. Venter: *The Sequence of the Human Genome*, *Science*, Bd. 291, 2000, S. 1304 – 1351.

[Wat95] M. S. Waterman: *Introduction to Computational Biology: Maps, Sequences and Genomes*, Chapman & Hall, London, 1995.

[Zha93] M. Q. Zhang, T. G. Marr: *A weight array method for splicing signal analysis*, *Computer Applications in Biosciences*, Bd. 9, Nr. 5, 1993, S. 499–509.

[Zha03] Z. Zhang, M. Gerstein: *Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements*, *Journal of Biology*, Bd. 2, Nr. 2, 2003, S. 11.