Problemstellungen der Bioinformatik

Proseminar im Grundstudium, Sommersemester 2007

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; [SM97]: Kap. 1.5 und 5; [Gus97]: Teil aus Kap. 16, [BB03]: 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 [SM97]: Kap. 4; [Cas]; [BB03], Kap. 8

4 Human Genome Project: Sequenzierung des menschlichen Genoms

Als Startpunkt: [Int01], [Mur02], [C.V00]

5 Phylogenetische Bäume *

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

Literatur [SM97]: Kap. 6; [Gus97]; [Wat95], [CC05], [HS03]

6 Einführung Massenspektrometrie *

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

7 Signalverarbeitung und Alignment von LC-MS-Daten *

- XCMS [SWO$^+$ss], [SAKG06]

8 Metabolitenidentifikation mit hochauflösenden Massenspektrometern

- MSMS Technik, Fragmentmuster, Substrukturen [Leh96]: Kap. 1.3
- Arabidopsis Profiling, MS/MS zur Identifizierung [RLDZ$^+$04]
- Exakte Masse, Isotopenmuster

9 Proteindatenbanken *

Literatur PDB: [BKW$^+$77]; [SLJ98]
SWISS-PROT : [ABSE04], `http://www.expasy.org/sprot/`

10 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:*
[BNP69]
Allgemein zur Vorhersageproblematik:
[KD97]; [LTZ96]
[CB00]: Teil aus Kap. 6, [Gla95]: Kap. 9.III

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

11 Protein-Ligand-Docking

Literatur FlexX `http://cartan.gmd.de/flexx/`; [JWG+97]

12 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]; [KPB+98]; [Hac99]; Das Affymetrix Benutzerhandbuch

Weitere Links `www.affymetrix.com`

13 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: [ESBB98]; [TSM+99]

14 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, BGL+99, MTMM04]

15 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 [SM97]: Kap. 3; [Gus97]; [Wat95]

16 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 [SM97]: Kap. 3.5; [Gus97]

17 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, ZM93, Sta84]

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

18 DNA-Computing

[BCJ+02, RL00]

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

Literatur

[ABSE04] Bairoch A., Boeckmann B., Ferro S., and Gasteiger E. Swiss-prot: Juggling between evolution and stability. *Briefings in Bioinformatics*, 5:39–55, 2004.

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

[BB03] Hans-Joachim Böckenhauer and Dirk Bongartz. *Algorithmische Grundlagen der Bioinformatik*. Teubner, 2003.

[BCJ+02] Ravinderjit S. Braich, Nickolas Chelyapov, Cliff Johnson, Paul W. Rothemund, and Leonard Adleman. Solution of a 20-variable 3-sat problem on a dna computer. *Science*, 296:499–502, 2002.

[BGL+99] M. Brown, W. Grundy, D. Lin, N. Christianini, C. Sugnet, M. Jr, and D. Haussler. Support vector machine classification of microarray gene expression data, 1999.

[BKW+77] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. (Jr.) Meyer, M. Brice, J. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The protein data bank: a computer–based archival file for macromolecular structures. *Journal of Molecular Biology*, 112:535–542, 1977.

[BNP69] W. J. Browne, A. C. T. North, and 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*, 42:65–86, 1969. Historisches Paper mit handgemachter Homologievorhersage, Drahtmodell und Stereophotos.

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

[Bur98] Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[Cas] D. Casey. Primer on molecular gentics. http://www.ornl.gov/sci/techresources/Human_Genome/publicat/primer/primer.pdf.

[CB00] Peter Clote and Rolf Backofen. *Computational Molecular Biology*. Wiley, 2000.

[CC05] Michael D. Crisp and Lyn G. Cook. Do early branching lineages signify ancestral traits? *Trends in Ecology & Evolution*, 20(3), 2005.

[C.V00] C.Venter. The sequence of the human genome. *Science*, 291:1304 – 1351, 2000.

[Dav04] Anthony N Davies. The new automated mass spectrometry deconvolution and identification system (amdis). *Spectroscopy Europe*, 10(3):22–26, 2004.

[ESBB98] Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, and David Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95:14863–14868, 1998.

[Gla95] Jay A. Glasel, editor. *Introduction to biophysical methods for protein and nucleic acid research*. Academic Press, 1995. Physikalische Beschreibung von Rntgenstrukturanalyse und Beschreibung von Faltungsvorhersage.

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

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

[HS03] Maureen Heymans and Ambuj K. Singh. Deriving phylogenetic trees from the similarity analysis of metabolic pathways. *Bioinformatics*, 19 Suppl., 2003.

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

[JWG+97] Gareth Jones, Peter Willett, Robert C. Glen, Andrew Leach, and Robin Taylor. Development and validation of a genetic algorithm for flexible docking. *Journal Molecular Biology*, 267:727–748, 1997.

[KD97] R. König and T. Dandekar. Computational methods for the prediction of protein folds. *Biochemica et Biophysika acta*, 1343(1):1, 1997.

[KPB+98] A. Kel, A. Ptitsyn, V. Babenko, S. Meier-Ewert, and H. Lehrach. A genetic algorithm for designing gene family-specific oligonucleotide sets used for hybridization: the g protein-coupled receptor protein superfamily. *Bioinformatics*, 14(3):259–270, 1998.

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

[LTZ96] T. Lengauer, R. Thiele, and R. Zimmer. Modellierung von proteinstrukturen. *Der GMD-Spiegel*, 2/3:14–18, 1996.

[MTMM04] Peter Meinicke, Maike Tech, Burkhard Morgenstern, and Rainer Merkl. Oligo kernels for datamining on biological sequences: a case study on prokaryotic translation initiation sites. *BMC Bioinformatics*, 5(1):169, 2004.

[Mur02] R. Mural. A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science*, 296(5573):1661 – 1671, May 2002.

[RL00] Adam J. Ruben and Laura F. Landweber. The past, present and future of molecular computing. *Nature Reviews Molecular Cell Biology*, 1:69–72, 2000.

[RLDZ+04] Edda von Roepenack-Lahaye, Thomas Degenkolb, Michael Zerjeski, Mathias Franz, Udo Roth, Ludger Wessjohann, Jürgen Schmidt, Dierk Scheel, and Stephan Clemens. Profiling of arabidopsis secondary metabolites by capillary liquid chromatography coupled to electrospray ionization quadrupole time-of-flight mass spectrometry. *Plant Physiology*, 134:548–559, February 2004.

[SAKG06] Dong-Guk Shin Dennis W. Hill Saira A. Kazmi, Samiran Ghosh and David F. Grant. Alignment of high resolution mass spectra: development of a heuristic approach for metabolomics. *Metabolomics*, 2(2):75–83, 2006.

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

[SLJ98] J. L. Sussman, D. Lin, and J. Jiang. Protein data bank (pdb): Database of three-dimensional structural information of biological macromolecules. *Acta Crystallographica Section D Biological Crystallography*, 6:1078–, 1998.

[SM97] Joao Setubal and Joao Meidanis. *Introduction to Computational Molecular Biology*. PWS Publishing, Boston, Mass., 1997.

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

[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*, 10:770–781, 1999.

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

[TSM+99] Pablo Tamayo, Donna Slonim, Jill Mesirov, Qing Zhu, Sutisak Kitareewan, Ethan Dmitrovsky, Eric S. Lander, and Todd R. Golub. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA*, 96(6):2907–2912, 1999.

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

[ZM93] M. Q. Zhang and T. G. Marr. A weight array method for splicing signal analysis. *Computer Applications in Biosciences*, 9(5):499–509, 1993.