Meilensteine der (Bio-)Informatik

Roland Wittler
roland ät cebitec.uni-bielefeld.de, U10-145

Sommersemester 2011

1 Administratives, Scheinkriterien

Do, 14-16 Uhr in S2-212

`http://gi.cebitec.uni-bielefeld.de/teaching/2011summer/meilensteine`

Vortrag:

- mindestens eine Woche vorher mit Veranstalter vorbesprechen
- 20 bis 25 Minuten pro Person
- Verfahren/Algorithmus, Beispiele, geschichtliche Einordnung, Relevanz, Paper kommentieren, Autor(en) vorstellen, ...

Reviewing:

- Optional: Nur wer ein Review schreibt, erhält auch eines.
- Wer ein (vernünftiges) Review schreibt bekommt "Rabatt" bei der Ausarbeitung.
- Review muss innerhalb von einer Woche angefertigt werden.
- Abgabe per Email (plain text) an Veranstalter

Ausarbeitung:

- Latex-Vorlage verwenden (siehe Webseite)
- mindestens 1500 Worte ($\approx$ 4–5 Seiten reiner Text)
- Abgabe: spätestens 3 Wochen nach Vortrag
- Wer am Reviewing teilnimmt: 1000 Worte, 1. Abgabe nach zwei Wochen, 2. Abgabe eine Woche nach Erhalt des Reviews.

2 Fragestellungen für Vortrag und Ausarbeitung

- Autoren:
  - Werdegang
  - Biologe oder Informatiker?
  - zum Zeitpunnkt des Papers: Alter, Karrierestand, Institut, ...
  - Verhältnis der Koautoren, Erstautor
  - Anekdoten?
  - Zeitstrahl?
- Paper:
  - Journal/Konferenz
  - Bio- oder Informatik-Paper?
  - Enstehung, Anekdoten?
  - Aufbau, Vorgehen, ...
  - Stil, Besonderheiten
  - Algorithmus wiedererkennbar?
- Inhalt:
  - Theorie, Algorithmus (wie bekannt vs. wie im Paper?)
  - Anwendung/Ergebnisse im Paper
  - historische Einordnung:
    Vorarbeiten, Weiterentwicklungen, Grundstein für ..., Nutzen/Anwendung heute, Zusammenhang zu anderen Papern, ...

3 Themen

Search, Sorting	
Quicksort	[Hoa61]
Boyer-Moore(-Horspool)	[BM77, Hor80]
Knuth-Morris-Pratt	[KMJP77]
Aho-Corasick	[AC75]
Suffix Trees	
Mc Creight	[McC76]
(Weiner	[Wei73])
Ukkonen	[Ukk95]
Manber-Myers	[MM90]
Comparison	
MUMmer	[DKF$^+$99, DPCS02, KPD$^+$04]
Needleman-Wunsch	[NW70]
Smith-Waterman	[SW81b, SW81a]
FASTA	[PL88]
BLAST, BLAST2	[AGM$^+$90], [AMS$^+$97]
Carrillo-Lipman	[CL88]
CLUSTAL-W	[THG94]
Mining, Compression	
Markov-Ketten	[Mar06]
GLIMMER	[SDKW98, DHK$^+$99]
Nussinov	[NJ80]
Lempel-Ziv	[ZL77]
Burrows-Wheeler	[BW94]
Biotechnology	
Sanger	[SNC77]
PCR	[SGS$^+$88]
Phylogeny	
Fitch-Hartigan	[Fit71, Har73]
Sankoff	[San75, SR75, SC83]
Maximum-Likelihood	[Fel81]
Graphs	
PQ-Trees	[BL76]
Euler-Tour	[Eul41, Eul56]

4 Hinweise

In diesem Abschnitt werden einige allgemeine Tipps und Regeln für die Vorbereitung und Durchführung eines Vortrags und das Schreiben einer Ausarbeitung gegeben. Die folgenden Auzählungen sollen als Leitfaden verstanden werden: Einerseits ohne Anspruch auf Vollständigkeit, andererseits kann es für einzelne Punkte auch Ausnahmefälle geben. Bitte nutzt diese 'Regeln' nicht nur für Eure eigene Arbeit, sondern auch als 'Checkliste' für konstruktives Feedback zu den Vorträgen und Ausarbeitungen der anderen Teilnehmer.

4.1 Vortrag

Aufbau

- Titelfolie, Motivation
- thematische Gliederung, "roter Faden"
- kurze, stichpunktartige Sätze
- nach Möglichkeit wenig Text, mehr Bild
- ggf. weitere Medien nutzen (z.B. Tafel)
- Animationen nur wenn inhaltlich sinnvoll
- "Zierrat" nur spärlich einsetzen
- Quellenangaben nicht vergessen (insbesondere von Bildern)

Durchführung

- den Vortrag proben, um ein gutes Zeitgefühl zu bekommen
- ca. ein bis zwei Minuten pro Folie
- Kontakt zum Publikum, Tempo überprüfen, eventuell Fragen an das Publikum stellen
- Klare und deutliche Aussprache
- angemessene Lautstärke (weder brüllen noch flüstern)
- gezielte Betonung/Wiederholung von wichtigem
- nicht ablesen oder auswendig lernen

4.2 Ausarbeitung

- Auf verständliche Gliederung und erkennbaren roten Faden achten.
- Keine extremen Bandwurmsätze konstruieren:
  - *eine* Idee bzw. *ein* Fakt pro Satz
  - *ein* Gedankengang pro Absatz
- Einfach, präzise, eindeutig und sachlich (nicht umgangssprachlich) formulieren.
- Jedes Kapitel vor dem ersten Unterabschnitt mit Einleitungs(ab)satz beginnen: Verbindung zwischen den Kapiteln, Motivation, Inhalt, Aufbau.
- Beispiele und Grafiken verwenden.
- *Alle* Bilder und Tabellen im Text referenzieren.
- Bei übernommenen Bildern Quellenangabe nicht vergessen.
- Tabellen und Bilder aussagekräftig beschriften. Üblicherweise: Tabellen*über*schriften und Bild*unter*schriften.
- Konsistenz: Begrifflichkeiten, Formatierung etc.
- Wikipedia u.ä. sind nicht zitierfähig.
- Plagiatismus hat Durchfallen zur Folge.
- Referenzen werden wie Fußnoten verwendet: “Smith und Waterman haben dies und das gemacht [1].” Nicht: “In [1] wurde dies und jenes gezeigt.”
- Rechtschreibung und Grammatik kontrollieren, auf Copy-Paste-Fehler achten.
- Im Deutschen wird im Passiv formuliert. “Man” sollte vermieden werden. Das dem englischen “we” entsprechende “wir” ist im Deutschen sehr unüblich.
- Grundsätzlich im Präsens schreiben — Ausnahmen sind durchaus sinnvoll.
- Neu eingeführte Begriffe bei ihrer ersten Verwendung kusiv setzen (`\emph{}`).
- Kursivierung sparsam/gezielt einsetzen, keine Unterstreichungen oder Fettdruck verwenden.
- Die Verwendung von Fußnoten ist nur selten wirklich sinnvoll.
- Abkürzungen vermeiden.

4.3 Review

- Zusammenfassung des Inhalts der Ausarbeitung (zwei bis vier Sätze)
- inhaltliche Kritik, auch positive Aspekte hervorheben, sachlich und begründet argumentieren
- technische Kritik, Rechtschreibung, Stil etc.
- Reviewer ist anonym

Literatur

[AC75] A. V. Aho and M. J. Corasick. Efficient string matching: An aid to bibliographic search. *Commun. ACM*, 18(6):333–340, 1975.

[AGM+90] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–410, 1990.

[AMS+97] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and psi-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.*, 25(17):3389–3402, 1997.

[BL76] K. S. Booth and G. S. Lueker. Testing for the consecutive ones property, interval graphs and graph planarity using *PQ*-tree algorithms. *J. Comput. Syst. Sci.*, 13(3):335–379, 1976.

[BM77] R. S. Boyer and J. S. Moore. A fast string searching algorithm. *Commun. ACM*, 20(10):762–772, 1977.

[BW94] M. Burrows and D. J. Wheeler. A block sorting lossless data compression algorithm. Technical Report TR 124, Digital Equipment Corporation, Palo Alto, CA, 1994.

[CL88] H. Carrillo and D. Lipman. The multiple sequence alignment problem in biology. *SIAM J. Appl. Math.*, 48(5):1073–1082, 1988.

[DHK+99] A. L. Delcher, D. Harmon, S. Kasif, O. White, and S. L. Salzberg. Improved microbial gene identification with Glimmer. *Nucleic Acids Res.*, 27(23):4636–4641, 1999.

[DKF+99] A. L. Delcher, S. Kasif, R. D. Fleischmann, J. Peterson, O. White, and S. L. Salzberg. Alignment of whole genomes. *Nucleic Acids Res.*, 27(11):2369–2376, 1999.

[DPCS02] A. L. Delcher, A. Phillippy, J. Carlton, and S. L. Salzberg. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.*, 30(11):2478–2483, 2002.

[Eul41] Leonhard Euler. Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, 8:128–140, 1741.

[Eul56] Leonhard Euler. The seven bridges of Königsberg. In James R. Newman, editor, *The World of Mathematics*, volume 1, pages 573–580. Simon and Schuster, New York, 1956.

[Fel81] J. Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.*, 17:368–376, 1981.

[Fit71] W. M. Fitch. Toward defining the course of evolution: Minimum change for a specific tree topology. *Syst. Zool.*, 20(4):406–416, 1971.

[Har73] J. A. Hartigan. Minimum mutation fits to a given tree. *Biometrics*, 29(1):53–65, 1973.

[Hoa61] C. A. R. Hoare. Algorithm 64: Quicksort. *Commun. ACM*, 4(7):321, 1961.

[Hor80] R. N. Horspool. Practical fast searching in strings. *Softw. Pract. Exper.*, 10(6):501–506, 1980.

[KMJP77] D. E. Knuth, J. H. Morris Jr., and V. R. Pratt. Fast pattern matching in strings. *SIAM J. Computing*, 6:323–350, 1977.

[KPD+04] S. Kurtz, A. Phillippy, A. L. Delcher, M. Smoot, Shumway M., C. Antonescu, and S. L. Salzberg. Versatile and open software for comparing large genomes. *Genome Biol.*, 5:R 12, 2004.

[Mar06] Andrei A. Markov. An example of statistical investigation of the text eugene onegin concerning the connection of samples in chains. *Science in Context*, 19.4:591–600, 2006. trans. David Link.

[McC76] E. M. McCreight. A space-economical suffix tree construction algorithm. *J. ACM*, 23(2):262–272, 1976.

[MM90] U. Manber and E. W. Myers. Suffix arrays: A new method for on-line string searches. In *Proc. First Annu. ACM-SIAM Symp. Discrete Algorithms, SODA 1990*, pages 319–327, 1990.

[NJ80] Rith Nussinov and Ann B. Jacobson. Fast algorithm for predicting the secondary structure of single-stranded rna. *Proc. Natl. Acad. Sci. USA*, 77(11):6309–6313, 1980.

[NW70] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48:443–453, 1970.

[PL88] W. R. Pearson and D. J. Lipman. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, 85(8):2444–2448, 1988.

[San75] D. Sankoff. Minimal mutation trees of sequences. *SIAM J. Appl. Math.*, 28(1):35–42, 1975.

[SC83] D. Sankoff and R. J. Cedergren. Simultaneous comparison of three or more sequences related by a tree. In D. Sankoff and J. B. Kruskal, editors, *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, chapter 9, pages 253–263. Addison-Wesley, Reading, MA, 1983.

[SDKW98] S. L. Salzberg, A. L. Delcher, S. Kasif, and O. White. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.*, 26(2):544–548, 1998.

[SGS+88] R.K. Saiki, D.H. Gelfand, S. Stoffel, S.J. Scharf, R. Higuchi, G.T. Horn, K.B. Mullis, and H.A. Erlich. Primer-directed enzymatic amplification of dna with a thermostable dna polymerase. *Science*, 239(4839):487–91, 1988.

[SNC77] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA*, 74:5463–5467, 1977.

[SR75] D. Sankoff and P. Rousseau. Locating the vertices of a steiner tree in an arbitrary metric space. *Math. Program.*, 9:240–246, 1975.

[SW81a] T. F. Smith and M. S. Waterman. Comparison of biosequences. *Adv. Appl. Math.*, 2:482–489, 1981.

[SW81b] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J. Mol. Biol.*, 147(1):195–197, 1981.

[THG94] J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22(22):4673–4680, 1994.

[Ukk95] E. Ukkonen. On-line construction of suffix trees. *Algorithmica*, 14(3):249–260, 1995.

[Wei73] P. Weiner. Linear pattern matching algorithms. In *Proc. of the 14th Annual IEEE Symposium on Switching and Automata Theory*, pages 1–11. IEEE Press, 1973.

[ZL77] J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Trans. Inf. Theory.*, 23(3):337–343, 1977.