

Prüfungsprotokoll

Kurs 01738 Grundlegende Algorithmen der Bio-Informatik

Datum: 08.03.2010

Prüfer: Dr. Rainer Merkl

Beisitzer: Hr. Zellner

Note: 1,0

Dauer: 25 min

Welche Verfahren des paarweisen Sequenzvergleichs gibt es?

Dynamische Programmierung: NW- und SW-Algorithmus, heuristische Verfahren wie FASTA und BLAST.

Welche Metrik findet hier Anwendung?

Levenshtein-Distanz: minimale Anzahl an Editieroperationen (Einfügen, Löschen, Ersetzen), um eine in die andere Sequenz umzuwandeln.

Wie kommt man dabei zu einer Distanz?

Kosten für die Operationen festlegen.

Wie findet die L.-Distanz Anwendung in den Algorithmen zum paarweisen Sequenzvergleich (Gleichung bei dynam. Progr.)?

Berechnung jeweils Minimum von Teilkosten + Kosten für Lücke bzw. Distanz beim Alignieren von zwei Symbolen. Einfügen/Löschen = Lücke in einer der beiden Sequenzen, Ersetzen = Alignieren zweier Symbole.

Unterschied NW und SW?

Globales vs. lokales Alignment.

(irgendwie kamen wir auch noch auf affine Kostenfunktionen zu sprechen, da weiß ich aber die Frage nicht mehr)

Wie erhält man bei SW das Alignment?

Höchsten Score in Matrix suchen und Backtracking bis zu einer 0 (die gehört dann nicht mehr zum Alignment).

Könnte man die Scores auch statt mit dynam. Progr. rekursiv berechnen bzw. warum macht man das nicht?

Weil man Teilergebnisse mehrfach berechnen müsste – zu hohe Laufzeit (dies wollte er hören), Rekursionstiefe

Woher erhält man die Scores?

Scoring-Matrizen (PAM, BLOSUM).

Wie werden die Scores bestimmt?

Bei BLOSUM aus Blocks-Datenbank, Betrachtung des gemeinsamen Auftretens von zwei Aminosäuren in einer Spalte: $\log(q(a_i, a_j) / (p(a_i) * p(a_j)))$.

Warum werden die Wahrscheinlichkeiten im Nenner multipliziert?

Zufälliges gemeinsames Auftreten von zwei unabhängigen Ereignissen.

Wo wird noch auf ähnliche Weise ein Score berechnet?

Bei Profilen: $\log(f(a_j, k) / f(a_j))$.

Woher kommt das, dass man die Scores so berechnet?

Neyman-Pearson-Lemma bzw. Bayesche Entscheidungstheorie: Entscheidung zugunsten der Alternativhypothese wenn Quotient größer als bestimmte Schwelle.

Welche Verfahren gibt es für phylogenetische Analysen?

Distanzbasierte (Neighbor-Joining-Algorithmus), Maximum-Parsimony- und Maximum-Likelihood-Methoden.

Was ist die einfachste?

Neighbor-Joining-Algorithmus.

Was muss man bei Maximum-Likelihood machen?

Bäume bestimmen und deren Wahrscheinlichkeit (bei bestimmter Beobachtung) berechnen.

Was ist leichter?

Wahrscheinlichkeit berechnen.

Welche Methode für Maximum-Likelihood wurde im Skript besprochen?

Quartett-Puzzle: Bestimmen der (n über 4) Quartette mit höchster Wahrscheinlichkeit, Beginnen mit einem, schrittweises Hinzufügen von weiteren Knoten (erhöhen der Scores von Kanten in die dieser nicht eingefügt werden soll mit Hilfe von Quartetten, in denen der Knoten vorkommt, Einfügen in die Kante mit niedrigstem Wert). Erstellen von verschiedenen Bäumen -> Konsensusbaum.

Was ist Bootstrapping?

Resampling-Methode um Topologie des Baumes zu überprüfen: zufällig Spalten aus MSA auswählen, Baum berechnen, schauen ob die gleichen Kanten auftauchen.

Warum sind MSAs wichtiger als paarweises Alignment?

Anforderungen an jede Position werden präziser beschrieben (Variation, Konserviertheit).

Welche Verfahren gibt es zum Erstellen von MSAs?

Progressives Alignment: zuerst zwei Sequenzen alignieren, dann schrittweise die anderen dazu. ClustalW (zuerst noch phylogenetischer Baum erstellt), T-Coffee

Warum ist T-Coffee besser?

Weil die Scores aus einer erweiterten Bibliothek kommen und man dadurch keine Scoring-Matrix braucht.

Wo finden neuronale Netze Anwendung?

Protein-Sekundärstruktur-Vorhersage (PHD-Algorithmus).

Wie wird dabei ein MSA genutzt?

Aus dem MSA wird ein Profil erstellt und dieses dann durch das n. N. ausgewertet.

Was ist der Grundbaustein eines n. N.?

Neuron bzw. Perzeptron: n Eingänge, jeweils gewichtet, ein Ausgang, Schwellenwertfunktion

Ist das alles festgelegt?

Schwellenwertfunktion ja, Gewichte nein (werden in Trainingsphase erlernt).

Welches Verfahren gibt es dafür?

Backpropagation-Algorithmus: schrittweises Verändern der Gewichte, Gradientenabstieg.

Was wird minimiert?

Mittleres Fehlerquadrat, partielle Ableitungen nach Gewichten werden berechnet.

Welche Gewichte werden am stärksten verändert?

Ich hab irgendwas mit in Richtung des Gradienten gesagt, er wollte hören: diejenigen die am meisten zum Fehler beitragen.

Markov-Modell für CpG-Inseln?

Ich hab zunächst das mit 8 Zuständen genannt, er wollte dann aber das leichte mit zwei Zuständen (CpG-Insel oder nicht) hören.

Was ist der Unterschied zwischen einer Markov-Kette und HMM?

Gemeint war das Lokalisationsproblem (im Gegensatz zum Diskriminationsproblem).

Wie macht man das beim Markov-Modell?

Fenster verschieben und jeweils dafür den Zustand vorhersagen.

Was ist das Problem dabei?

Das Fenster hat eine feste Größe, man kann die CpG-Insel nicht genau lokalisieren.

Was ist ein HMM?

Stochastischer Prozess, der eine Beobachtung und mit ihr verschränkten Pfad im Zustandsgraph erzeugt.

Was ist hidden?

Pfad ist verborgen, man sieht nur die Beobachtung.

Wie kann man den Pfad bestimmen?

Viterbi-Pfad.

Wie bekommt man den?

Über Viterbi-Variablen, Berechnung im Viterbi-Algorithmus, mit Hilfe von Maximierung über die Viterbi-Variablen der Vorgänger-Zustände (genaue Berechnung wollte er nicht wissen)

Wozu gehört der Algorithmus?

Dynamische Programmierung.

Profil-HMMs: was macht man damit?

Proteindomänen modellieren, Zugehörigkeit zu Protein-Familie testen.

Wie kommt man vom MSA zu Profil-HMM?

Emissionswahrscheinlichkeiten aus Trainingsdaten (Match-Zustände: Häufigkeiten in MSA-Spalten, Insertions-Zustände: Hintergrundwahrscheinlichkeiten, Deletionszustände haben keine Emission).

Woher weiß man wie viele Match-Zustände man braucht?

MSA: wenn in einer Spalte mehr Zeichen als bestimmter Schwellenwert, gibt es einen Match-Zustand.

An die genauen Fragen konnte ich mich ehrlich gesagt nicht mehr so genau erinnern, aber der ungefähre Ablauf der Prüfung sollte deutlich geworden sein. Die Prüfungsatmosphäre war locker und die Fragen sehr fair. Es wurde nur das grobe Verständnis und keine fiesen Details abgefragt. Die Prüfung fand in der Uni Regensburg statt. Herr Merkl war nett und hat geholfen und nachgefragt, wenn ich nicht direkt wusste, worauf er hinauswill.

Datum: 29.03.2010

Prüfart: Regensburg

Prüfer: Dr. Merkl

Zeit: keine 20 min

1. Warum werden überhaupt Sequenzen verglichen?
2. Welche Rückschlüsse kann man aus einer ähnlichen Struktur/ Funktion schließen?
3. Welche Distanzbegriffe kennen Sie?
4. Wie wird bei der Hamming-Distanz der Unterschied gemessen? Wie kommt die Distanz zustande?
5. Wie wird bei der Levensthein-Distanz der Unterschied gemessen, welche Editieroperationen gibt es?
5. Wie lautet die Formel zur Levensthein-Distanz? Malen sie bitte ein Dotplot auf und zeigen sie wie die Formel die Matrize befüllt.
6. Wie werden Lücken dargestellt, in welcher Sequenz von den hier aufgemalten ist eine Lücke vorhanden?
7. Wie erhält man dann die Lösungssequenz? Was ist das Backtracking und wie funktioniert das?
8. Wie wird die Matrize initialisiert beim NW Algo und wie wird sie dann berechnet und was möchte ich hier maximieren? Wie heißt der andere große Vertreter?
9. Wie wird die Matrize beim SW Algo initialisiert? Wieso wird hier gerade auf Null beschränkt?
10. Was ist der Nachteil beider Verfahren?
11. Wie sind PAM und BLOSUM Score-Matrizen aufgebaut? Wie und auf welcher Basis wird bei BLOSUM der Score berechnet?
12. Schreiben sie die Formel zum logg-odds score auf, und erklären sie jeweils beide Terme. Was bedeutet ein positiver Score, was ein negativer Score im Zusammenhang mit der Formel?
13. Wie wurden die Scores wohl in der BLOSUM Datenbank bestimmt? Was bedeutet BLOSUM 100, Was BLOSUM 45? Was ist der Unterschied zum PAM Score?
11. Wie kann man die schlechte Laufzeit von n^2 umgehen? Wie lauten die Verfahren dazu?
12. Erklären sie den Ablauf beim BLAST Verfahren. Was bedeutet HSP in dem Zusammenhang? Was bedeutet die Abkürzung HSP?
13. Wie werden HSPs berechnet, welcher räumliche Abstand ist hier gemeint? Können jegliche hits zum HSP erweitert werden?

Soweit reicht meine Erinnerung noch, leider musste ich feststellen dass hier doch ein tiefgreifendes Verständnis und Faktenwissen vorhanden sein musste um hier eine gute Note abzugreifen.

Es ärgerte mich auch, dass hier ausschließlich ein großer Themenblock (Distanzen und Scores, das waren EA 2 ein bißchen EA 3) abgefragt wurde, keine Spur von HMM, MSA, Stammbäume, Neuronale Netze oder genetische Algorithmen usw.

Prüfungsprotokoll

Kurs 1738 Grundlegende Algorithmen der Bio-Informatik*

Datum: 26.03.2009

Prüfer: Dr. Rainer Merkl

Note: 1,3

Dauer: 30 Min

Die folgende Liste gibt den ungefähren Verlauf der Fragen in der Prüfung wieder. Natürlich alles ohne Anspruch auf Vollständigkeit und Korrektheit :-)

Q1: Welche Methoden gibt es um Stammbäume zu erzeugen?

A1: Distanzbasierte und Gruppierung nach Charaktereigenschaften (Parsimony)

Q2: Welchen Algorithmus kennen Sie für die distanzbasierte Methode?

A2: Neighborhood-Joining

Q3: Was ist das Prinzip des Neighborhood-Joining?

A3: Ermitteln der Sequenzen mit den größten Ähnlichkeiten; Bildung eines neuen Knotens; Iteratives Hinzufügen der anderen Knoten.

Q4: Wie funktionieren die Parsimony-Algorithmen?

A4: Anhand des Beispiels zur Datenkompression erklärt.

Q5: Welche anderen Formen des Sequenzalignments kennen Sie?

A5: Multiples Sequenzalignment

Q6: Welchen Vorteil haben die MSAs gegenüber den paarweisen?

A6: Die Anforderungen an den einzelnen Positionen werden besser beschrieben.

Q7: Welchen Distanzbegriff kennen Sie?

A7: Levenstein (spricht sich im Übrigen "Löwenstein"; nur falls sich jemand wie ich während der Prüfung fragt, welcher Algorithmus das sein soll :-))

Q8: Was verbirgt sich hinter diesem Distanzbegriff?

A8: Min. Anzahl an Editieroperationen (Einfügen, Löschen, Ersetzen) um eine Sequenz in eine andere zu überführen.

Q9: Noch mal zu den phylogenetischen Bäumen. Wie würden Sie dort die Ähnlichkeit berechnen?

A9: Da es sich um eine evolutionäre Betrachtung handelt => paarweiser Vergleich mit dem Needleman-Wunsch-Algorithmus (globales Alignment).

Q10: Welche Verbindung gibt es zwischen HMM und MSA?

A10: Profil-HMM

Q11: Können Sie ein HMM mit zwei Zuständen für das CpG-Insel-Problem aufzeichnen?

A11: Ja. (Hab ich dann auch gemacht :-))

Q12: Wie werden die Wahrscheinlichkeiten ermittelt?

A12: Trainingsmenge und Häufigkeitsanalyse (Baum-Welch-Parameterschätzung musste nicht erklärt werden :-))

Q13: Wie berechnet sich die Wahrscheinlichkeit für eine Insertion?

A13: Aus den Hintergrundwahrscheinlichkeiten

Q14: Wie wird Wissen der Anwendungsdomäne in die Algorithmen zum paarweisen Sequenzvergleich gebracht?

A14: Durch Scoringmatritzen

Q15: Wie werden diese generiert?

A15: Unterschiedlich für PAM und BLOSUM

Q16: Bitte am Beispiel BLOSUM

Q16: Substitutionswahrscheinlichkeiten in den einzelnen Positionen.

Q17: Wie werden diese berechnet?

A17: $\log(q(a_i, a_j)/p(a_i)p(a_j))$: Die Formel wurde anschließen weiter ausgeführt.

Q18: Warum werden die Wahrscheinlichkeiten des Nenners multipliziert?

A18: Die Wahrscheinlichkeiten sind voneinander unabhängig (zufälliges Modell).

Q??: Welche Algorithmen erwarten MSAs als Eingabe?

A??: Zur Profilgenerierung, Verstärken einer Query bei Abfrage gegen eine Datenbank, NN bei PHD

Q??: Welche Eingabe erwartet PHD?

A??: Query-Sequenz => Generierung des MSAs => Generierung des Profils

Q??: Was sind Suffix-Bäume und was geben sie an?

A??: Gemeinsame Suffixe von Sequenzen.

Die Prüfung fand in Regensburg bei Herrn Dr. Merkl statt. Die Prüfungsatmosphäre war locker und sehr freundlich. Es wurden zu den einzelnen Themengebieten die Prinzipien angefragt und die grundsätzliche Funktionsweise und falls zutreffend ihre Anwendung in der Bio-Informatik. Formeln wurden bis auf eine Ausnahme nicht verlangt.

Falls man bei einer Fragestellung die Antwort nicht sofort weiß, hilft Herr Merkl einem durch Tipps weiter.

Alles in allem kann ich diesen Kurs nur weiterempfehlen, auch wenn man anfangs ein wenig zweifelt, ob die biologischen Vorkenntnisse reichen.

Dieser Kurs hat seinen Schwerpunkt eindeutig auf den Algorithmen, die auch wenn man sie vielleicht nicht so in der täglichen Informatik nutzt, sehr interessante Konzepte und Ideen aufweisen.

Prüfungsprotokoll
Kurs 1738 Bioinformatik

An diesem Termin haben zwei Prüfungen stattgefunden. Da sich der Ablauf bei beiden Prüfungen nur bei den letzten Fragen unterscheid, haben wir die beide Prüfungen im einem Protokoll zusammengefasst.

Datum: 17.05.2004
Prüfer: Dr. R. Merkl
Beisitzer: Keller (Diplomand)
Ergebnis: 1,3 bzw. 1,0
Prüfungsdauer: Jeweils 25 Min.

- Was ist das Prinzip beim Sequenzvergleich; warum ist es überhaupt interessant Sequenzen zu vergleichen?
 - Sequenzen modellieren biologische Strukturen. Wenn Sequenzen eine hinreichend große Ähnlichkeit zueinander aufweisen, kann geschlossen werden, dass eine ähnliche Struktur und damit auch eine ähnliche Funktion vorliegt.
- Gilt hierzu auch die Umkehrung, und ab wann kann man den von einer hinreichenden Ähnlichkeit sprechen?
 - Nein, die Umkehrung gilt nicht! Ab mehr als 30 – 35 % identischer Residuen kann man von hinreichender Ähnlichkeit sprechen.
- Wie kann man Ähnlichkeit quantifizieren?
 - Über Distanzen. Distanzen und Ähnlichkeit sind Dual zueinander. Je geringer die Distanz zwischen 2 Sequenzen, desto größer ist die Ähnlichkeit.
- Wie wird das bei der Levenshteindistanz ausgedrückt?
 - Die Levenshteindistanz gibt die minimale Anzahl an Editieroperationen an, die notwendig ist um eine Sequenz A in eine andere Sequenz B überzuführen. (Einfügen, Löschen und Ersetzen von Symbolen). Die Formel wurde nicht verlangt.
- Wie wurde diese Idee beim NW-Algorithmus umgesetzt?
 - Hier wird nicht die Distanz sondern der Score berechnet, der der Ähnlichkeit entspricht. (Ein hoher Score entspricht hoher Ähnlichkeit und geringer Distanz). Der Score wird maximiert! (Minimum Kosten, Einfügen von Lücken, affine Kostenfunktion)
- Es gibt noch einen weiteren solchen Algorithmus. Wie heißt der, und was sind die Unterschiede zum NW-Algorithmus?
 - Es gibt noch den Smith-Waterman-Algorithmus. Der berechnet im Unterschied zum NW nicht den globalen Score, sondern einen lokalen. Dazu muss der Score nach unten mit 0 beschränkt werden.
- Warum sind lokale Alignments denn so interessant; wo liegt hier der Vorteil?
 - Weil sie sich besser zum Vergleich von Proteindomänen eignen. Proteindomänen sind die kleinsten Einheiten mit einer definierten und unabhängig gefalteten Struktur. Sie besitzen individuelle Funktionen innerhalb eines Proteins. Sie bestehen meist aus 50 bis 150 Residuen und bilden die bekannten Sekundärstrukturelemente α -Helix, β -Strang, -Faltblatt, ... aus.
- Was sind die Probleme bei diesen beiden Algorithmen?
 - Die Laufzeit! Sie liegt in $O(n^2)$.
- Das ist ein Problem, wenn man eine Sequenz gegen eine ganze Datenbank vergleichen will. Wie kann man das Lösen?

- Mit heuristischen Methoden zum Sequenzvergleich. Diese verwenden Preprozessingschritte um ähnliche Teilsequenzen mit zu identifizieren und zu indizieren.
- Was ist in diesem Zusammenhang zu beachten, wenn an eine Datenbank Abfragen, die auf verschiedenen Scoringsystemen beruhen, gestellt werden sollen?
 - Für jedes zu verwendende Scoringsystem muss ein eigener Index existieren, da das Scoringsystem großen Einfluss auf die Scoreberechnung hat. Z.B. BLAST unterstützt einige verschiedene Substitutionsmatrizen.
- Erklären sie den Algorithmus der in BLAST verwendet wird.
 - 1. *Preprozessing*: Erstellen einer Liste aller w-mers die einen gewissen Score überschreiten.
 - 2. *Lokalisierung der hits*: Bestimmen der Positionen der gemeinsamen Vorkommen der w-mers in den Vergleichssequenzen.
 - 3. *Bestimmung der HSPs* (High-Scoring Segment-Pais): Paare von hits die auf der selben Diagonale liegen und deren Abstand kleiner als ein vorab festgelegter Schwellwert A ist. Beginn und Ende der HSPs sind so gewählt, dass sowohl eine Verlängerung als auch eine Verkürzung ihren Score verringert.
 - 4. *Erweiterung mit Lücken*: Aus den HSPs die einen Schellwert überschreiten wird dasjenige mit höchstem Score gewählt. Davon ausgehend wird mittels Dyn. Prog. das Alignment in beide Richtungen erweitert. Dabei werden nur solche Zellen betrachtet für die der errechnete Score im Vergleich zum bisherigen maximalen Score um weniger als X sinkt.
- Themenwechsel; Neuronale Netze: Wie ist ein Perzeptron aufgebaut?
 - Aufgezeichnet: Gewichtung der Eingänge, Summierung, Schwellwertfunktion
- Wie wird nun ein Netz mit gegebener Architektur trainiert?
 - Durch Schrittweise und gerichtete Modifikation der Gewichte. Gradientenabstieg.
- Nun zu genetischen Algorithmen: was sind die Probleme die hier auftreten können?
 - Habe die Schwierigkeit eine adäquate Kodierung für ein Problem zu finden; Erwähnt. Das war aber nicht gemeint. Gesucht war der Umstand, dass keine Garantie besteht, das globale Minimum zu erreichen.
- Dieses Problem besteht auch im Zusammenhang mit NN. Wie kann man bei NN und GA diesem Problem begegnen?
 - NN: Paralleles Training ausgehend von verschiedenen initialen Gewichten; Wenn sich in mehreren Netzen ähnliche Gewichte einstellen, kann man davon ausgehen, dass man das globale Minimum gefunden hat.
 - GA: Das jeweils beste und das jeweils schlechteste Individuum werden unverändert in die nächste Generation übernommen.
- Letzte Frage: HMMs: Wie funktionieren HMMs zur Bestimmung von MSAs?
 - Den erw. Zustandsgraph eines Profil-HMM gezeichnet. Die Match-, Insertion-, Deletion-Zustände und deren Emissionen erklärt (2 verschränkte stochastische Prozesse).
- Wie kann man denn die Emissionswahrscheinlichkeiten bestimmen?
 - Bei Match-Zuständen aus den positionsabhängigen Häufigkeiten der Aminosäuren.
 - Bei Insertion-Zuständen aus den Hintergrundwahrscheinlichkeiten.
 - bei Deletion-Zuständen wird mit Wahrscheinlichkeit 1 ein „-“ emittiert.
- Was ist ein Viterbi-Pfad?
 - Der wahrscheinlichste Pfad auf dem eine konkret gegebene Beobachtung emittiert wird.

Die Prüfung fand ausnahmsweise am Institut für Mikrobiologie und Genetik an der Universität Göttingen statt.

Wir waren etwas zu früh am Prüfungsort. Dr Merkl zeigte uns das Sequenzierlabor und erklärte anhand der verschiedenen Stationen den Ablauf bei der Sequenzierung und Annotation.

Das Prüfungsgespräch selbst lief unter dem Motto „keep it simple“: Es wurden keine Formeln gefragt, die Fragen zielten auf guten Überblick. Bei Algorithmen ist das Konzept (Warum und wie macht man das? Was ist die biologische Begründung für diese Vorgangsweise?) und die damit verbundenen Problematiken wichtig. Wenn man bei Fragen Probleme hatte, versuchte Dr. Merkl zu unterstützen. Sobald er das Gefühl hatte, dass ein Thema gut verstanden war, wechselte er zu einem Anderen.

Dr. Merkl ist ein sehr sympathischer und angenehmer Prüfer. Bemerkenswert ist, dass er sogar seinen Urlaub unterbrochen hat, um uns unseren Terminwunsch zu erfüllen. Die Prüfung verlief wie ein Gespräch. Zwischendurch erklärte er Zusammenhänge ergänzend zum Buch aus der Sicht der Praxis.

Viel Erfolg zu Euren Prüfungen.