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 Levenshteindisanz 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 Swith-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 verschiede 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 verbunden 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.