Erinnerungsprotokoll

Mündliche Prüfung Kurs 01738 „Grundlegende Algorithmen der Bioinformatik“
30 Minuten, 13.05.2015, Prof. Merkl

- Zuerst zu Taxonomie. Was ist das und warum macht man das?
  Finden einer Vererbungshierarchie und Erzeugung phylogenetischer Bäume
- Was sind phylogenetische Bäume und was für Probleme hat die Informatik damit?
  Darstellung von Vererbungshierarchien. Problem 1 ist, den Baum zu erzeugen. Dafür gibt es Heuristiken. Problem 2 ist anschließend zu überprüfen, ob der Baum die Beobachtungen wiedergibt.
- Wie kann man nun einen Baum erzeugen?
  Neighbour Joining Algorithmus oder Quartett Puzzle. Mit Erklärung, wie die funktionieren.
- Wie überprüft man den Baum anschließend?
  Maximum Likelihood. Wie wahrscheinlich ist es, dass der Baum stimmt?
- Wie funktioniert Likelihood?
  Anhand der Formel erklären.
- Was ist der Unterschied zwischen überwachtem und unüberwachtem, maschinellem Lernen?
  Überprüfen des Lernvortschrittes anhand der Testdaten.
- Was ist ein neuronales Netz?
  Ausgehend vom Neuron den Aufbau geschildert. Stichworte: Prezeptron, Schwellwertfunktion, Eingaben und Ausgaben, Gewichtungen.
- Wie trainiert man, wie kontrolliert man?
  Stichworte: Bis die Veränderung des Fehlers unter einen Schwellwert fällt, Trainingsmenge, Leave one out und Crossvalidierung.
- Wie vermeidet man lokale Minima?
  Häufiger versuchen und wenn ein Ergebnis oft vorkommt, ist die Chance groß, dass es stimmt. Wichtig: Vergleichen der Ergebnisse. Beim Neuronalen Netz sollten nicht die Gewichtungen verglichen werden. Die können durchaus unterschiedlich für das gleiche Ergebnis sein.
- Was sind die Voraussetzungen für dynamisches Programmieren?
  Teilergebnisse in O(n) und Zwischenspeichern. Zugriff auf die Zwischenergebnisse in O(1)
- Was ist ein Profil-HMM? Wie wird es erzeugt?
  Analog Kurstext schildern.
- Welche Wahrscheinlichkeiten werden dafür benötigt?
  Übergangswahrscheinlichkeiten und Emmissionswahrscheinlichkeiten. Auch erklären, wie die ermittelt werden.
- Wofür werden die Profil-HMM verwendet?
  Um bei einer neuen Sequenz zu beurteilen, ob die zu dem Profil gehört.
- Wie beurteilt man das?
  Viterbi Pfad berechnen. Erklärt, wie das gemacht wird. Stichworte: Viterbi Variablen. Wie wahrscheinlich ist es, dass von dem HMM die neue Sequenz emittiert wird?
- Wovon hängen Viterbi Variablen ab?
  Nur vom Vorgangszustand im HMM.

Gut fand ich, dass recht konkrete Fragen gestellt wurden, auf die ich antworten konnte. Ist mir lieber als ein Vortrag. Hin und wieder waren wir uns nicht so ganz einig über das Ziel einer Frage, aber das ließ sich immer lösen. „Das stimmt zwar, darauf wollte ich aber nicht hinaus“. Ein weiterer Eindruck war, dass Prof. Merkl sehr genau auf die Zeit achtete und immer wieder mal auch hart einen Bruch machte, um noch ein weiteres Thema ansprechen zu können. Gut so.

Prüfungsprotokoll Kurs 01738
„Grundlegende Algorithmen der Bioinformatik“
April 2013

Wie schon von anderen beschrieben, ist Herr Prof. Merkl sehr freundlich und geduldig und hilft einem während der Prüfung auf die Sprünge, wenn man mal feststeckt. Er schien auch nicht gelangweilt, wenn man nur „das Übliche“ erzählte, was er sicher schon x-mal gehört hat, sondern hörte aufmerksam zu.

Noch ein Wort zu den weiteren Umständen der Prüfung: Zunächst fiel es mir nicht leicht, sein Büro an der Uni Regensburg zu finden (das kann aber auch an mir selbst gelegen haben ;)), evtl. sollte man also ein paar Extraminuten für die Suche einplanen. Schmierzettel und Stift liegen bereit und man darf gerne Gebrauch davon machen.

Herr Prof. Merkl meinte mit einem Augenzwinkern, er habe sich ein paar neue Fragen ausgedacht und tatsächlich war meine Prüfung etwas anders als jene, die man sonst in den Protokollen nachlesen konnte. Das brachte mich zum Teil sehr ins Schwimmen, weil ich darauf nicht so gut vorbereitet war. Es folgt nun ein Gedächtnisprotokoll der Prüfung, wie immer ohne Gewähr für Vollständigkeit und Korrektheit. ;) Bei meinen Antworten sind nur die (hoffentlich) wesentlichen Stichpunkte genannt.

Prof. Merkl: **Wir haben uns ja besonders mit biologischen Sequenzen beschäftigt. Welche Datenbanken kennen Sie denn dafür?**
[Schon diese Frage brachte mich etwas in Verlegenheit, da ich mir die entsprechenden Abschnitte (Kap. 3) nicht genau angesehen hatte.]

Ich: BLOCKS (Blöcke von Aminosäuresequenzen, MSAs), Prosite (Charakteristische, kurze Sequenzteile, Motive)

Prof. Merkl: Das ist richtig, aber es sind ja schon relativ „high-level“ Datenbanken (sekundäre Datenbanken), **was gibt es denn da an ganz einfachen Sammlungen?**

Ich: PFAM (Proteinfamilien), EMBL (DNA-Sequenzen)

Prof. Merkl: Das ist auch noch zu abgehoben. Ich wollte auf „SWISS-PROT“ hinaus.

Prof. Merkl: **Beim Vergleich von Sequenzen, wie wird deren Ähnlichkeit bewertet und woher kommt das?**

Ich: Distanz und Ähnlichkeit (zueinander dual), man verwendet fast immer Scores, die Ähnlichkeit beschreiben. Scoring-Matrizen für Paare von Aminosäuren, Inhalte werden auch durch biologisches Fachwissen geprägt.

Prof. Merkl: **Welche zwei großen Arten von Algorithmen gibt es denn zum Sequenzvergleich?**

Ich: [War nicht ganz sicher, worauf er hinaus wollte.]
Meinen Sie damit Algorithmen der dynamischen Programmierung gegenüber heuristischen Algorithmen oder globales vs. lokales Alignment?

Prof. Merkl: Letzteres.

Ich: Beim globalen Alignment kann es vorkommen, dass eine Sequenz über die andere „verschmiert“ wird. Nicht gut z. B. zum Auffinden einzelner Proteindomänen. Lokale Alignments versuchen stattdessen, die Anfragesequenz möglichst geschlossen wiederzufinden.

Prof. Merkl: **Und wie wird solch ein lokales Alignment erreicht?**

Ich: Einfügen von Null in die Vergleichsfunktion → jede Position kann theoretisch wieder Anfang eines neuen lokalen Alignments sein.

Prof. Merkl: **OK, und woher kommt diese Null?**

Ich: [War sehr unsicher, weiß leider nicht mehr genau, was ich geantwortet habe. Es schien jedenfalls nicht ganz das zu sein, was er hören wollte.]

Prof. Merkl: **Warum sind Scores denn besser als Distanzen?**

Ich: Geben die Ähnlichkeit der beteiligten Residuen wieder, enthalten also auch viel Fachwissen. Scores können negativ werden, Distanzen (zumindest in der üblichen Definition) nicht. In der Regel ist der Erwartungswert des Scores negativ, sodass eine erste Aussage über die biologische Relevanz des Vergleichs bereits anhand des Vorzeichens getroffen werden kann.

Prof. Merkl: **Und wie werden die Scores bei den genannten Algorithmen berechnet bzw. eingesetzt?**

Ich: [Wieder recht unsicher, da ich dachte, alles schon erwähnt zu haben.] Vergleich über Einträge in der Scoring-Matrix, nochmal negativen Erwartungswert erwähnt.

Prof. Merkl: **Ich meine, was geschieht mit den Scores der einzelnen Positionen?**

Ich: [Noch immer unsicher:] Sie werden aufsummiert.
[Das war es wohl tatsächlich, worauf er hinauswollte – ich hatte vermutlich etwas zu kompliziert gedacht.]

Prof. Merkl: Genau! **Und besteht dabei ein Zusammenhang zwischen den Scores an den einzelnen Positionen?**

Ich: Nein – jede Position wird für sich ausgewertet. Einzige Ausnahme ist ggf. die affine Kostenfunktion (Kosten für Verlängerung einer Lücke werden eingesetzt, wenn zuvor eine Lücke eingeführt wurde.)

Prof. Merkl: **Noch einmal zur Herkunft der Scores, können Sie dazu noch etwas Genaueres sagen? Da gab es so ein Lemma…**

Ich: Neymann-Pearson-Lemma. Quotient der Wahrscheinlichkeiten muss größer sein als ein bestimmter Schwellenwert c, dann entscheide für Alternativhypothese. [Ich hatte Abb. 4.6 auf Seite 86 des Kurstexts vor Augen und versucht, sie zu beschreiben.] Wahrscheinlichkeitsverteilungen, c liegt „in der Mitte“, Änderung bzw. Verschiebung beeinflusst Wahrscheinlichkeit von Fehlern erster oder zweiter Art.

Prof. Merkl: **Wie nennt man diesen Quotienten genau?**

Ich: [Das fiel mir leider gerade nicht ein.]

Prof. Merkl: Es ist der Likelihood-Quotient. **Was ist denn der Unterschied zwischen „Likelihood“ und „Wahrscheinlichkeit“?**

Ich: [Hier bin ich wieder geschwommen…] Die Likelihood wird benötigt, um nach der Bayes’schen Entscheidungstheorie eine A-Posteriori-Wahrscheinlichkeit in eine A-Priori-Wahrscheinlichkeit umzuwandeln. [Ich hatte die Bayes’sche Formel im Kopf.]

Prof. Merkl: Ich wollte darauf hinaus, dass es sich um eine Wahrscheinlichkeit unter gegebener Beobachtung handelt.

Prof. Merkl: **Sequenz-Logos, was ist das?**

Ich: Geben Wahrscheinlichkeit des Vorkommens von Symbolen in Sequenzen positionsabhängig an. Je höher die Schriftgröße eines Symbols, desto wahrscheinlicher ist sein Vorkommen an

der jeweiligen Position.

Prof. Merkl: **Wie genau wird die Höhe des Symbols berechnet?**

Ich: Anhand der Wahrscheinlichkeit, das Symbol in der Spalte vorzufinden. Ist es z. B. strikt konserviert, wird nur dieses Symbol in der größten Schriftgröße angezeigt.

Prof. Merkl: **Genauer bitte (Tipp: Shannon)**

Ich: [Versucht, die Gleichung aus Abschnitt 10.6, S. 189 des Basistextes aufschreiben und mich daran weiterzuhangeln.] Im Zähler: Posititionsspezifische Wahrscheinlichkeit für eine bestimmte Aminosäure, im Zähler insgesamt betrachtete Wahrscheinlichkeit… Ergebnis ist also sowohl von der Position als auch von der Gesamthäufigkeit des Symbols abhängig.

Prof. Merkl: OK. Konkret wollte ich auf den spezifischen Informationsgehalt eines Symbols hinaus. (Seltene Aminosäuren haben einen höheren als häufigere.)

Prof. Merkl: Themenwechsel: **Phylogenetische Analysen. Welche Methoden gibt es da?**

Ich: Phaenetische vs. Kladistische Methoden. Erstere setzen auf Distanzen, letztere auf Maximum Parsimony oder Maximum Likelihood.

Prof. Merkl: **Direkt zu Maximum Likelihood, wie ist hier das genaue Vorgehen?**

Ich: Im Grund wird davon ausgegangen, dass der phylogenetische Baum bereits existiert. Die Wahrscheinlichkeit, dass er von den gegebenen Sequenzen erzeugt wird, kann dann anhand eines Modells geschätzt werden. Es gibt Heuristiken, um einen Baum zu kreieren, der dann untersucht werden kann.

Prof. Merkl: **Woher kommt das Modell für die Schätzung?**

Ich: [Etwas unsicher.] Es kann aus biologischen Hintergrundwahrscheinlichkeiten und dem „Pool" der Eingabesequenzen abgeleitet werden.

Prof. Merkl: **Sie sagten, es gebe Heuristiken für die Konstruktion des Baumes. Bitte erläutern.**

Ich: Quartett-Puzzle. Erstellung aller möglichen Quartette aus Eingabesequenzen, „möglichst sinnvolle" Einordnung zufällig ausgewählter Quartette in den wachsenden Baum. Mehrfache Wiederholung zur Ableitung des Konsensus-Baumes.

Prof. Merkl: **Was ist „Bootstrapping"?**

Ich: Zufälliges Auswählen von Spalten aus dem Eingabe-MSA der Baumkonstruktion, Erstellung eines neuen Baumes basierend auf diesen Spalten, Erhöhung des Gewichtes aller gemeinsamen Kanten mit dem erzeugten Baum um 1. Mehrfache Wiederholung. Ziel: Ausschließen, dass der Baum nur von wenigen Spalten abhängig ist.

Prof. Merkl: **Warum möchte man das ausschließen?**

Ich: Baum ist nur dann biologisch sinnvoll, wenn er möglichst die kompletten Sequenzen berücksichtigt. Hängt er nur von wenigen Spalten/Residuen ab, ist er wahrscheinlich nicht korrekt.

Prof. Merkl: Letztes Thema, **Profil-HMMs: Wie werden diese gebildet?**

Ich: Anhand einer Heuristik können für die einzelnen Spalten Insertions-, Match- und Deletionszustände gebildet werden.

Prof. Merkl: **Wie kommt man zu den Wahrscheinlichkeiten?**

Ich: Deletionszustände emittieren immer „-", Insertionszustände: Hintergrund-

wahrscheinlichkeiten, Match-Zustände: Aus spaltenweisen Häufigkeiten. Übergangswahrscheinlichkeiten aus Schätzern. Oftmals ist es sinnvoll, Pseudocounts mit einzuführen.

Prof. Merkl: Hier hätte mir noch das Stichwort „Baum-Welch-Algorithmus“ zur Parameterschätzung gefehlt.

Warum sind Profil-HMMs genauer als der Vergleich zweier Sequenzen mit den anderen Methoden?

Ich: Profil-HMMs bauen auf MSAs auf, daher von vornherein genauere Definition der Erwartung in einzelnen Spalten. Weiterhin sind sie genau auf die Problemstellung (Pool der Eingabesequenzen) abgestimmt.

Prof. Merkl: **Sind die Bewertungen der einzelnen Positionen hier auch unabhängig voneinander?**

Ich: [Jetzt fiel der Groschen ;)] Nein. Die Übergangswahrscheinlichkeiten von einem Zustand in den nächsten drücken die Abhängigkeit der Positionen voneinander aus. Dadurch wird die Genauigkeit natürlich auch erhöht.

Bewertet wurde die Prüfung letzten Endes mit 1,3. Das fand ich gerade auch angesichts meiner Unsicherheiten, besonders am Anfang, äußerst fair. Grundsätzlich habe ich immer versucht, möglichst viel von meinem Wissen herüberzubringen, auch wenn mir gerade die richtige Antwort nicht einfiel oder ich nicht wusste, worauf Prof. Merkl hinaus wollte. Ich schätze mal, das hat geholfen. ☺

Fazit: Prüfung und Benotung waren sehr fair, die Atmosphäre locker und sehr freundlich. Obwohl ich die Prüfung persönlich bisher die schwierigste fand (weil der Stoff doch eher theoretischer Natur ist), würde ich Kurs und Prüfer weiterempfehlen.

Prüfungsprotokoll
Kurs 01738 Grundlegende Algorithmen der Bioinformatik
Datum: 08.03.2013
Prüfer: Prof. Dr. Rainer Merkl
Note: 1,3
Dauer: 30 min

Was sind Ontologien?
Flexible Alternative zu starren hierarchischen Klassifikationssystemen; Begriffe und ihre Beziehungen; Verknüpfung von Typen mithilfe von Relationen (wichtigste Relation ist *is_a*); Gen-Ontologie: hierarchisches Annotationssystem mit streng kontrolliertem Vokabular. Beschreibung von Genprodukten wie z.B. molekulare Funktion, biologischer Prozess u. zelluläres Kompartiment.

Wie werden Scoring-Matrizen gefüllt?
Durch den Vergleich zweier Hypothesen (die Sequenzen sind miteinander verwandt bzw. nicht verwandt) mit der Neyman-Pearson-Methode. Zur Entscheidung wird der Quotient $\frac{p(x\,|H_1)}{p(x\,|H_0)} > c$ gebildet. Entscheidung zugunsten der Alternativhypothese wenn das Verhältnis größer als der Schwellenwert c ist.

Wie wird der Schwellenwert bestimmt?
Ich nenne den Schwellenwert jetzt mal γ. Die Schwelle γ wird anwendungsspezifisch gewählt, indem man die Anzahl falsch positiver oder falsch negativer Vorhersagen festlegt.

Wie genau kommt man dabei zum Schwellenwert? Da gibt es doch so eine Kennlinie.
Mithilfe einer ROC-Kurve.

Zeichnen Sie bitte eine ROC-Kurve und erläutern Sie anschließend die Vorgehensweise.
Eine ROC-Kurve gezeichnet, Achsen mit TPR und FPR beschriftet und erwähnt, dass der Flächeninhalt unterhalb der Kurve die Qualität des Klassifikators spezifiziert. Anschließend gezeigt, wie man z.B. durch Auswahl eines Wertes für TPR die Schwelle γ und den Wert FPR erhält.

Wie wird eine ROC-Kurve aufgenommen?
Zusammenstellen eines Testdatensatzes, Festlegen der kritischen Grenzen γ_{min}, γ_{max} und des Inkrements γ_{inc}. Beginne bei $\gamma = \gamma_{min}$ und wiederhole bis $\gamma \geq \gamma_{max}$ die folgenden Schritte:
Berechne TPR und FPR in Abhängigkeit vom Testdatensatz und von γ
$\gamma = \gamma + \gamma_{inc}$
Ausgabe: in jeder Runde TPR, FPR und γ

Was ist der Unterschied zwischen der ROC-Kurve und den Klassifikationsverfahren?
Gemeint war, dass bei der ROC-Kurve ein Testdatensatz bzw. eine Trainingsmenge benötigt wird, während bei den Klassifikationsverfahren Objekte miteinander verglichen werden.

Was für Klassifikationsverfahren gibt es?
Einfaches iteratives Clusterverfahren, k-mean-Clusterverfahren, agglomeratives hierarchisches Clusterverfahren, Nächster-Nachbar-Klassifikation.

Welche Methoden/Verfahren zur Vorhersage der Sekundärstruktur kennen Sie?
Zur Vorhersage der Protein-Sekundärstruktur die Chou-Fasman-Methode und PHD. Vorhersage der RNA-Sekundärstruktur durch Energieminimierung oder unter Verwendung eines genetischen Algorithmus, z.B. im Programm STAR.

Bleiben wir bei der Protein-Sekundärstrukturvorhersage. Wie ist PHD aufgebaut?
Architektur des PHD-Systems: Programm zur Berechnung eines MSAs, gefolgt von zwei hintereinandergeschalteten neuronalen Netzen mit jeweils einem hidden layer. Zum Schluss ein Mittelwertbildner.

Weshalb werden zwei Netzwerke hintereinandergeschaltet?
In der Trainingsphase werden die Elemente der Trainingsmenge in zufälliger Reihenfolge ausgewählt. Daher kann das erste neuronale Netz keine Korrelation zwischen aufeinanderfolgenden Residuen „erlernen". Das zweite neuronale Netz (Struktur → Struktur) ist jedoch aufgrund der präsentierten Daten hierzu in der Lage.

Woraus ergeben sich die Werte für den Mittelwertbildner?
Die Trainingsmenge muss aus Paaren (Sequenz, bekannte Sekundärstruktur) bestehen. In diesen Datensätzen sind die drei Sekundärstrukturelemente α-Helix, β-Faltblatt und Sonstiges (L) nicht gleich häufig vertreten. Daher gibt es zwei Arten von Training: balanced und unbalanced Training. Für je eine Version beider Netze wurde balanced und unbalanced Training eingesetzt. Daraus ergeben sich 2 x 2 Kombinationen von neuronalen Netzen und hieraus resultieren die vier Werte aus denen als maximaler Mittelwert die Sekundärstrukturvorhersage abgeleitet wird.

Warum sind MSAs wichtig?
Weil die Anforderungen an die einzelnen Positionen präziser beschrieben werden. Z.B. tritt die Konserviertheit von Residuen aus dem statistischen Rauschen hervor und die Variabilität lässt sich untersuchen.

Beim paarweisen Sequenzvergleich unterscheiden wir zwischen lokalen und globalen Alignments. Wie kommt man vom globalen zum lokalen Alignment bzw. worauf muss geachtet werden?
Beim Berechnen eines lokalen Alignments muss der Score nach unten hin mit null beschränkt werden. Jede Position hat somit dieselbe Chance der Anfang eines neuen globalen Alignments zu werden.

Weshalb wählt man die null und nicht irgendeinen anderen Wert wie z.B. π oder e?
Das war etwas kniffliger. Am Ende lief es auf die folgende Formel hinaus:

$$\sum \frac{q(as_i, as_j)}{p(as_i)p(as_j)} = log \prod \frac{q(as_i, as_j)}{p(as_i)p(as_j)} = log\ \frac{P(A,B \mid V)}{P(A,B \mid Z)}$$

Wobei Prof. Merkl den linken Ausdruck als Hilfestellung gab. Auf den Rest musste man selber kommen. Der rechte Ausdruck entspricht dem logarithmierten Quotienten der Likelihoodwerte für die Hypothesen V bzw. Z. Dieser Term ist gleich null wenn $P(A,B \mid V) = P(A,B \mid Z)$. Daher wird die null als Schwellenwert festgelegt.

Kommen wir zu den Hidden-Markov-Modellen. Da haben wir die Profil-HMM kennengelernt. Wie kommt man vom MSA zum Profil-HMM?

1. Bestimmen der Match-Zustände. Die Länge n des Profil-HMMs, d.h. die Anzahl der Dreierblöcke von Deletions-, Insertions- und Match-Zuständen mit gleichem Index, ist gleich der Anzahl dieser Match-Zustände. Jeder Spalte des MSAs, in der mehr Buchstaben vorkommen als ein bestimmter Schwellenwert, wird ein Match-Zustand zugeordnet. Damit hat man die Topologie des Profil-HMMs.
2. Bestimmen der Emissionswahrscheinlichkeiten
Bei Match-Zuständen aus den positionsabhängigen Häufigkeiten im MSA (Profil des MSAs).
Bei Insertions-Zuständen aus den Hintergrundswahrscheinlichkeiten.
Bei Deletions-Zuständen wird mit Wahrscheinlichkeit 1 ein „–“ emittiert.

Können Sie den Begriff der Hintergrundswahrscheinlichkeiten genauer erläutern? Was kann man sich darunter vorstellen?
Das sind die relativen Häufigkeiten, mit denen die Buchstaben im MSA vorkommen.

So kann man das auch ausdrücken. Was fehlt nun noch?
Die Übergangswahrscheinlichkeiten. Diese erhält man über die Maximum-Likelihood-Schätzer: Bestimmen eines Pfades für jede Zeile des MSAs. Anwenden der Gleichungen der Maximum-Likelihood-Schätzer für die Übergangs- und Emissionswahrscheinlichkeiten.

Können Sie noch ein wenig mehr zu den Maximum-Likelihood-Schätzern erzählen?
Maximum-Likelihood-Schätzer sind problematisch wenn die Trainingsmenge klein ist. Dann kann es sein, dass einige der Größen, die man zu deren Berechnung benötigt, gleich null sind, obwohl das für die zu schätzenden Wahrscheinlichkeiten nicht der Fall ist. Die Lösung dieses Problems ist das Einbringen von Pseudocounts, die man den Größen hinzuaddiert als Ausdruck von A-priori-Wissen. Die einfachste Art Pseudocounts einzusetzen ist die Laplacesche Regel: zu jeder Häufigkeit wird ein hinzuaddiert.

Kennen Sie noch eine andere Art Pseudocounts einzubringen?
Dirichlet-Schätzer.

Was ist ein HMM?
Ein diskreter stochastischer Prozess, der eine Beobachtung und einen mit ihr verschränkten Pfad erzeugt.

HMM und Markov Ketten sind für eine besonders wichtige Fragestellung der Bioinformatik von Bedeutung. Können Sie diese benennen?
Das Problem der Lokalisierung von CpG-Inseln.

Wie wird das Lokalisationsproblem mit der Markov Kette gelöst?
Ein Fenster der Länge l wird über die Eingabesequenz verschoben und für jede mögliche Anfangsposition k wird der Score berechnet. (Die Formel wollte er nicht wissen.) Teilwörter mit positivem Score sind potentielle CpG-Inseln. Problem: die Länge von CpG-Inseln ist unbekannt, während die Fenstergröße fix ist. Ist das Fenster zu groß, sind CpG-Inseln nur Teilwörter des Fensters und der Score eventuell zu klein. Ist das Fenster zu klein, unterscheiden sich das Plus- und Minus-Modell nicht hinreichend um eine gute Trennung zu ermöglichen.

Wie erhält man die Übergangswahrscheinlichkeiten innerhalb und außerhalb von CpG-Inseln?
Durch eine Häufigkeitsanalyse auf einer hinreichend großen, unabhängigen Trainingsmenge von CpG-Inseln bzw. von Strings, die keine CpG-Inseln sind.

Welche Markov-Modelle für CpG-Inseln kennen Sie?
Das Modell mit 9 bzw. 3 Zuständen (den Startzustand mitgezählt).

Was ist das Besondere am Modell mit 9 Zuständen?
Die Emissionswahrscheinlichkeiten sind entartet. D.h. jeder Zustand emittiert mit Wahrscheinlichkeit 1 seinen eigenen Buchstaben.

Nennen Sie eine Methode zur Erzeugung von Stammbäumen/phylogenetischen Bäumen.
Maximum-Likelihood-Methoden. Gesucht wird ein phylogenetischer Baum, dessen Wahrscheinlichkeit bei gegebenen Taxa maximal ist.

Was ist die Schwierigkeit dabei?
Mithilfe eines Modells für die Mutationsvorgänge lässt sich die Wahrscheinlichkeit für das Auftreten der betrachteten Sequenzen anhand eines bestimmten, gegebenen Baumes berechnen. Schwieriger ist es, die Topologie des wahrscheinlichsten Baumes zu bestimmen.

Welches Verfahren kennen Sie dafür?
Eine Heuristik dafür ist das Verfahren der Quartett-Puzzles.

Erläutern Sie bitte die Vorgehensweise beim Quartett-Puzzle.
1. Schritt: Für alle $\binom{n}{4}$ Quartette von Taxa werden die drei möglichen Maximum-Likelihood-Bäume konstruiert. Die Bäume mit den höchsten Maximum-Likelihood-Werten sind optimal und werden im Puzzleschritt verwendet.
Puzzle-Schritt: Die Reihenfolge der Eingabe wird randomisiert, z.B. A, B, C, D, E, …
Beginn mit dem Maximum-Likelihood-Baum von (A, B, C, D)
Schrittweises Einfügen der restlichen Taxa. Unter Verwendung der optimalen Quartette, die das Taxon E enthalten und deren Nachbarschaftsrelationen werden die Scores der Kanten erhöht, in denen das Taxon E nicht eingesetzt werden soll.
Konsensus-Baum: → erster temporärer Baum konstruiert
Wiederhole Puzzle-Schritt mehrere Male
Ableiten eines Konsensus-Baumes aus den resultierenden temporären Bäumen

Für die Richtigkeit und Vollständigkeit der Fragen und Antworten übernehme ich keine Gewähr. Ferner sei darauf hingewiesen, dass das Studieren der Prüfungsprotokolle allein zur Vorbereitung auf die Prüfung nicht ausreicht! Ansonsten kann ich mich nur meinen Vorrednern anschließen. Prof. Merkl ist sehr nett und hilft einem mit Tipps weiter, wenn man die Antwort nicht sofort weiß bzw. formuliert die Frage anders, falls man nicht versteht worauf er hinaus möchte. Die Fragen waren allesamt fair und tiefgreifende Details wurden nicht verlangt.

Viel Erfolg bei Euren Prüfungen!

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: $\log(q(as_i,as_j)/(p(as_i)*p(as_j)))$.

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: $\log(f(as_i, k)/f(as_i))$.

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, Deletionszustande 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.

Kurs 1738 Grundlagen der Bioinformatik

Datum: 29.03.2010
Prüfort: 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 intialisiert 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 initalisiert? 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 postiver 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: Levensthein (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(ai, aj)/p(ai)p(aj)): 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 Vorkentnnisse 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 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.