Andrew Torda
Björn Hansen
Iryna Bondarenko

Zentrum für Bioinformatik

Übung zur Vorlesung „Bioinformatik“

Sommersemester 2015

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

18.05.2015

Übung 4: Revision

Beispielfragen zur Klausur im Modul „Bioinformatik“ (erste Semesterhälfte)

Anmerkungen:

- Die Prüfungsfragen der Klausur werden auf Deutsch gestellt.
- Sie basieren sowohl auf der Vorlesung als auch auf der Übung.
- Auf den folgenden Seiten finden Sie typische Fragen, wie sie in der Klausur gestellt werden könnten. Dies ist aber kein Fragenkatalog, sondern nur eine kleine Sammlung von möglichen Prüfungsfragen.

Frage 1:

Sie haben zwei DNA-Sequenzen, für die ein Alignment angefertigt werden soll:
ACGTCCTTCATT und GTCTCATG

a) Sie haben das folgende Bewertungsschema:

- Übereinstimmung (match): +1
- Mismatch: -1
- Kosten für das Öffnen einer Lücke (gap opening costs): -1
- Kosten für das Erweitern einer Lücke (gap extension costs): -1

Scheiben Sie das beste lokale Alignment dieser beiden Sequenzen auf.

b) Sie haben das folgende Bewertungsschema:

- Übereinstimmung (match): +1
- Mismatch: 0
- Kosten für das Öffnen einer Lücke (gap opening costs): -10
- Kosten für das Erweitern einer Lücke (gap extension costs): 0

Scheiben Sie das beste lokale Alignment dieser beiden Sequenzen auf.

Frage 2:

Sie haben eine Score-Matrix für ein Paar von DNA-Sequenzen berechnet.
Nach der Traceback-Berechnung erhalten Sie folgendes Ergebnis:

	A	C	A	C	C	T	T	A
C								
C								
A								
T								
C								
C								
A								
A								

Schreiben Sie das zugehörige Sequenzalignment inklusive aller Lücken (gaps) an den richtigen Positionen auf.

Frage 3:

Bei Protein-Alignments werden nicht nur Matches und Missmatches betrachtet, sondern es wird auch die Ähnlichkeit von Aminosäuren berücksichtigt. Wie werden diese Ähnlichkeiten repräsentiert?

Frage 4:

Sie berechnen Alignments von Proteinsequenzen unter Verwendung der folgenden Substitutionsmatrix:

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

a) Kosten für das Öffnen einer Lücke (gap opening costs): -8

Kosten für das Erweitern einer Lücke (gap widening/extension costs): -1

Sie erhalten das folgende Alignment:

AACDQRST
A-CD-RST

Wie lautet der Score-Wert dieses Alignments?

b) Angenommen Sie bekämen das folgende Alignment:

AACDQRST
A-CD--ST

Wie lautete jetzt der Score-Wert?

c)

AACDQRST
A-CD-SST

Wie lautet der Score-Wert für dieses Alignment?

Frage 5:

Für Berechnung von Proteinsequenz-Alignments können viele verschiedene Substitutionsmatrizen verwendet werden. Warum könnte eine bestimmte Substitutionsmatrix gegenüber einer anderen bevorzugt werden?

Frage 6:

Oft kommt die Substitutionsmatrix BLOSUM zum Einsatz.
Skizzieren Sie die nötigen Schritte zur Berechnung der Werte in dieser Matrix.

Frage 7:

a) Worin besteht der Vorteil eines Needleman-Wunsch-Alignments gegenüber einem „seeded"-Alignment" (z.B. einem BLAST-Alignment)?

b) Was ist der Vorteil einer „seeded"-Methode" wie BLAST gegenüber dem Needleman-Wunsch-Verfahren?

c) Nennen Sie eine Anwendung, bei welcher Sie Alignments bevorzugt mit BLAST und nicht mit dem Needleman-Wunsch-Verfahren berechnen würden.

d) Nennen Sie eine Anwendung, bei welcher Sie das eher langsame Needleman-Wunsch-Verfahren einem BLAST-ähnlichen Verfahren vorziehen würden.

Frage 8:

a) Was ist der Unterschied zwischen einer wiederholt durchgeführten BLAST-Suche (PSI-BLAST) und einer gewöhnlichen BLAST-Suche?

b) Was ist der Vorteil von PSI-BLAST gegenüber einer gewöhnlichen BLAST-Suche?

Frage 9:

Sie haben zwei entfernt verwandte Proteine mit schwacher Sequenzähnlichkeit. Sie kennen sowohl die zugehörigen Protein- als auch die DNA-Sequenzen. Wieso würden Sie ein besseres Alignment erwarten, wenn Sie die Proteinsequenzen für das Alignment verwenden?

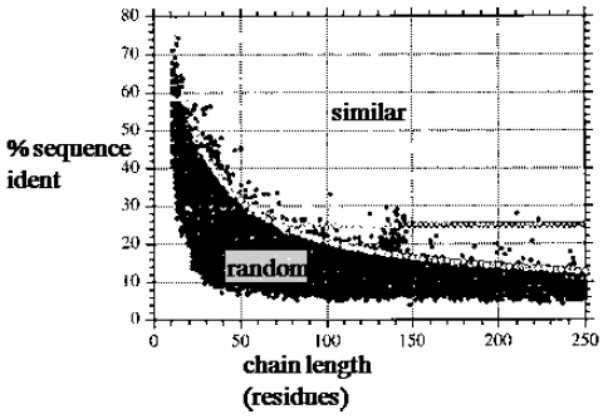
Frage 10:

a) Ich habe ein Programm erstellt, welches DNA-Zufallssequenzen erzeugt. Für paarweise Alignments solcher Sequenzen erwarte ich eine durchschnittliche Sequenzübereinstimmung von ca. 25%. Bei der Berechnung stelle ich aber fest, dass die Sequenzübereinstimmung meistens höher als 25% ist. Was könnte der Grund dafür sein?

b) Wenn die Sequenzen nicht zufällig erzeugt, sondern stattdessen zufällig aus einer Datenbank für biologische Sequenzen ausgewählt werden, ist die beobachtete Sequenzähnlichkeit sogar noch größer. Wie ist das zu erklären?

Frage 11:

a) Wenn wir für alle strukturell unterschiedlichen Paare von Proteinen aus einer großen Menge von Proteinen (z.B. der PDB) die Kettenlänge gegen die Sequenzähnlichkeit auftragen, dann entsteht ein Diagramm wie das Folgende:



Welche Aussagen lassen sich anhand dieser Abbildung treffen?

b) Ich habe zwei Proteine mit einer Sequenzübereinstimmung von 20% und möchte wissen, wie signifikant diese Sequenzähnlichkeit ist. Welche andere Information benötigte ich noch, um diese Frage beantworten zu können?

Frage 12

a) Was zeichnet ein korrektes Alignment aus?
Geben Sie eine biologisch und eine algorithmisch/mathematisch begründete Antwort!

b) Häufig ist das Finden des korrekten Alignments erschwert. Nennen und erläutern Sie drei Gründe dafür!

Frage 13

Ich möchte ein multiples Sequenzalignment für N_{seq} Sequenzen berechnen. Wie viele paarweise Alignments muss ich hierzu berechnen?

Frage 14

In einem multiplen Sequenz-Alignment wollen sie einen Score-Wert maximieren:

$$score = \sum_{b \neq \grave{a}}^{N_{seq}} \sum_{a=1}^{N_{seq}} S_{a,b}$$

Was bedeutet dieser Score-Wert in Worten?

Frage 15

Ich möchte einen Leitbaum (guide tree) für ein multiples Sequenzalignment erstellen. Wodurch wird die Reihenfolge festgelegt, in welcher die Knoten des Baumes verknüpft werden?

Frage 16

Ich habe 3 Sequenzen: A, B und C. Die Sequenzen B und C sind beide mit A verwandt, aber für ein Alignment zwischen B und C erhalte ich keinen guten Alignment-Score-Wert. Was könnte die Ursache dafür sein? Zeichnen Sie ein Diagramm, falls Ihnen das die Erklärung erleichtert.

Frage 17

Ich habe ein multiples Sequenz-Alignment berechnet. Ich möchte herausfinden, welche Positionen im Alignment konserviert sind und welche variieren. Deshalb erstelle ich eine Abbildung, in welcher die Konservierung bzw. die Variabilität gegen die Sequenzposition aufgetragen wird.

Sie erinnern Sich sicher an die Formel $S = \sum_{i=1}^{N_{states}} p_i \ln p_i$.

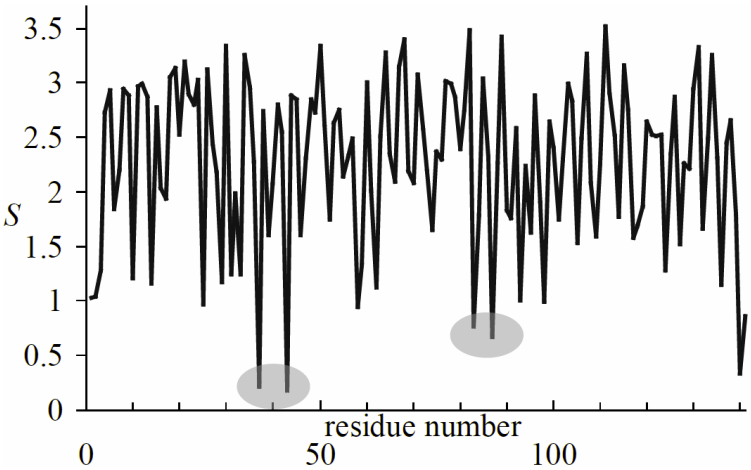
Welche Bedeutung hat p_i in dieser Formel?

Frage 18

Wie würden Sie vorgehen, um in einem multiplen Sequenz-Alignment potentiell katalytisch wichtige Seitenketten im aktiven Zentrum zu identifizieren?
Warum könnten Sie enttäuscht werden?

Frage 19:

Anhand eines multiplen Sequenz-Alignments habe ich Variabilität/Konservierung als eine Funktion der Sequenzposition berechnet.



a) Die Aminosäure-Reste 37 und 43 scheinen sehr gut konserviert zu sein.
Warum könnten diese beiden Reste wichtig sein?

b) Einige Reste in der Abbildung sind nicht besonders gut konserviert. Ich behaupte deshalb, dass diese Reste für die Funktion des Proteins unbedeutend sind.
Warum könnte ich damit falsch liegen?

Frage 20:

Ich habe ein Sequenzalignment von 400 Tyrosin-Kinasen berechnet und stelle fest, dass nur wenige Sequenzpositionen im Verlauf der Evolution konserviert wurden. Was könnte ich an meiner Auswertung ändern, um mehr konservierte Reste zu beobachten?

Frage 21:

Wir haben eine Familie von Sequenzen sowie alle zugehörigen paarweisen Alignments. Für zwei Sequenzen dieser Familie können wir die Anzahl von Unterschieden (Mutationen) zählen und dann mit folgender Formel den Anteil an mutierten Aminosäure-Resten berechnen:

$$p_{mut} = \frac{N_{diff}}{N_{length}}$$

Warum ist dies keine gute Methode?

Frage 22:

Ich möchte überlagerte (alinierte) DNA-Sequenzen für die Erstellung eines phylogenetischen Baumes verwenden. Nennen Sie zwei Gründe dafür, dass die Verzweigungen des Baumes unter Umständen nicht verlässlich sind.

Frage 23:

Ich habe einen phylogenetischen Baum mit einer Neighbor-Joining-Methode erstellt. Wie könnte ich vorgehen, um die Verlässlichkeit des berechneten Baumes zu überprüfen?

Frage 24:

Man nimmt an, dass sich Protein-Sequenzen im Verlauf der Evolution schneller ändern als Proteinstrukturen. Was könnte eine evolutionäre Ursache dafür sein?

Frage 25:

Sie haben ein bakterielles Protein mit unbekannter Funktion. Sie haben eine Knockout-Mutante hergestellt, aber ohne das entsprechende Gen stirbt das Bakterium sofort. Wie würden Sie vorgehen, um etwas über die Funktion dieses Proteins herauszufinden? Nach welcher Art von Information würden Sie suchen?

Frage 26

Warum ist die PDB keine gute Stichprobe für natürliche Proteine?

Frage 27

Welche Darstellungsform würden Sie in *Chimera* für ein Protein auswählen, wenn Sie die Sekundärstruktur sehen möchten?

Frage 28

Die Standardabweichung ist ein Maß für die Streuung einer Wertemenge $(x_1, x_2, ..., x_N)$ um deren Mittelwert $\overline{x}$. Mit der folgenden Formel lässt sich die Standardabweichung s berechnen:

$$s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

Der RMSD-Wert (Root-mean-square deviation) als Maß für den Unterschied zwischen zwei Molekülstrukturen lässt sich mit einer sehr ähnlichen Formel berechnen. Was würde in dieser Formel anstelle von $x_i - \overline{x}$ auftauchen?

Frage 29

Warum ist es schwierig einen RMSD-Wert zu berechnen, wenn zwei Strukturen unterschiedlich groß sind?

Andrew Torda
Björn Hansen
Iryna Bondarenko

Zentrum für Bioinformatik

Übung zur Vorlesung „Bioinformatik“

Sommersemester 2015

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

18.05.2015

Übung 4: Revision

Beispielfragen zur Klausur im Modul „Bioinformatik“ (erste Semesterhälfte)

Anmerkungen:

- Die Prüfungsfragen der Klausur werden auf Deutsch gestellt.
- Sie basieren sowohl auf der Vorlesung als auch auf der Übung.
- Auf den folgenden Seiten finden Sie typische Fragen, wie sie in der Klausur gestellt werden könnten. Dies ist aber kein Fragenkatalog, sondern nur eine kleine Sammlung von möglichen Prüfungsfragen.

Frage 1:

Sie haben zwei DNA-Sequenzen, für die ein Alignment angefertigt werden soll:
ACGTCCTTCATT und GTCTCATG

a) Sie haben das folgende Bewertungsschema:

- Übereinstimmung (match): +1
- Mismatch: -1
- Kosten für das Öffnen einer Lücke (gap opening costs): -1
- Kosten für das Erweitern einer Lücke (gap extension costs): -1

Scheiben Sie das beste lokale Alignment dieser beiden Sequenzen auf.

b) Sie haben das folgende Bewertungsschema:

- Übereinstimmung (match): +1
- Mismatch: 0
- Kosten für das Öffnen einer Lücke (gap opening costs): -10
- Kosten für das Erweitern einer Lücke (gap extension costs): 0

Scheiben Sie das beste lokale Alignment dieser beiden Sequenzen auf.

Frage 2:

Sie haben eine Score-Matrix für ein Paar von DNA-Sequenzen berechnet.
Nach der Traceback-Berechnung erhalten Sie folgendes Ergebnis:

	A	C	A	C	C	T	T	A
C								
C								
A								
T								
C								
C								
A								
A								

Schreiben Sie das zugehörige Sequenzalignment inklusive aller Lücken (gaps) an den richtigen Positionen auf.

Frage 3:

Bei Protein-Alignments werden nicht nur Matches und Missmatches betrachtet, sondern es wird auch die Ähnlichkeit von Aminosäuren berücksichtigt. Wie werden diese Ähnlichkeiten repräsentiert?

Frage 4:

Sie berechnen Alignments von Proteinsequenzen unter Verwendung der folgenden Substitutionsmatrix:

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

a) Kosten für das Öffnen einer Lücke (gap opening costs): -8

Kosten für das Erweitern einer Lücke (gap widening/extension costs): -1

Sie erhalten das folgende Alignment:

AACDQRST
A-CD-RST

Wie lautet der Score-Wert dieses Alignments?

b) Angenommen Sie bekämen das folgende Alignment:

AACDQRST
A-CD--ST

Wie lautete jetzt der Score-Wert?

c)

AACDQRST
A-CD-SST

Wie lautet der Score-Wert für dieses Alignment?

Frage 5:

Für Berechnung von Proteinsequenz-Alignments können viele verschiedene Substitutionsmatrizen verwendet werden. Warum könnte eine bestimmte Substitutionsmatrix gegenüber einer anderen bevorzugt werden?

Frage 6:

Oft kommt die Substitutionsmatrix BLOSUM zum Einsatz.
Skizzieren Sie die nötigen Schritte zur Berechnung der Werte in dieser Matrix.

Frage 7:

a) Worin besteht der Vorteil eines Needleman-Wunsch-Alignments gegenüber einem „seeded"-Alignment" (z.B. einem BLAST-Alignment)?

b) Was ist der Vorteil einer „seeded"-Methode" wie BLAST gegenüber dem Needleman-Wunsch-Verfahren?

c) Nennen Sie eine Anwendung, bei welcher Sie Alignments bevorzugt mit BLAST und nicht mit dem Needleman-Wunsch-Verfahren berechnen würden.

d) Nennen Sie eine Anwendung, bei welcher Sie das eher langsame Needleman-Wunsch-Verfahren einem BLAST-ähnlichen Verfahren vorziehen würden.

Frage 8:

a) Was ist der Unterschied zwischen einer wiederholt durchgeführten BLAST-Suche (PSI-BLAST) und einer gewöhnlichen BLAST-Suche?

b) Was ist der Vorteil von PSI-BLAST gegenüber einer gewöhnlichen BLAST-Suche?

Frage 9:

Sie haben zwei entfernt verwandte Proteine mit schwacher Sequenzähnlichkeit. Sie kennen sowohl die zugehörigen Protein- als auch die DNA-Sequenzen. Wieso würden Sie ein besseres Alignment erwarten, wenn Sie die Proteinsequenzen für das Alignment verwenden?

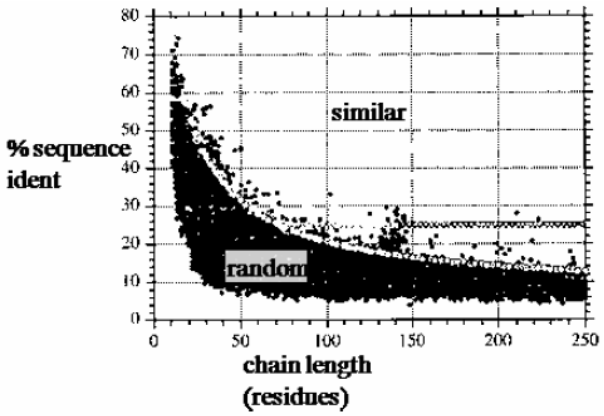
Frage 10:

a) Ich habe ein Programm erstellt, welches DNA-Zufallssequenzen erzeugt. Für paarweise Alignments solcher Sequenzen erwarte ich eine durchschnittliche Sequenzübereinstimmung von ca. 25%. Bei der Berechnung stelle ich aber fest, dass die Sequenzübereinstimmung meistens höher als 25% ist. Was könnte der Grund dafür sein?

b) Wenn die Sequenzen nicht zufällig erzeugt, sondern stattdessen zufällig aus einer Datenbank für biologische Sequenzen ausgewählt werden, ist die beobachtete Sequenzähnlichkeit sogar noch größer. Wie ist das zu erklären?

Frage 11:

a) Wenn wir für alle strukturell unterschiedlichen Paare von Proteinen aus einer großen Menge von Proteinen (z.B. der PDB) die Kettenlänge gegen die Sequenzähnlichkeit auftragen, dann entsteht ein Diagramm wie das Folgende:



Welche Aussagen lassen sich anhand dieser Abbildung treffen?

b) Ich habe zwei Proteine mit einer Sequenzübereinstimmung von 20% und möchte wissen, wie signifikant diese Sequenzähnlichkeit ist. Welche andere Information benötigte ich noch, um diese Frage beantworten zu können?

Frage 12

a) Was zeichnet ein korrektes Alignment aus?
Geben Sie eine biologisch und eine algorithmisch/mathematisch begründete Antwort!

b) Häufig ist das Finden des korrekten Alignments erschwert. Nennen und erläutern Sie drei Gründe dafür!

Frage 13

Ich möchte ein multiples Sequenzalignment für N_{seq} Sequenzen berechnen. Wie viele paarweise Alignments muss ich hierzu berechnen?

Frage 14

In einem multiplen Sequenz-Alignment wollen sie einen Score-Wert maximieren:

$$score = \sum_{b \neq \grave{a}}^{N_{seq}} \sum_{a=1}^{N_{seq}} S_{a,b}$$

Was bedeutet dieser Score-Wert in Worten?

Frage 15

Ich möchte einen Leitbaum (guide tree) für ein multiples Sequenzalignment erstellen. Wodurch wird die Reihenfolge festgelegt, in welcher die Knoten des Baumes verknüpft werden?

Frage 16

Ich habe 3 Sequenzen: A, B und C. Die Sequenzen B und C sind beide mit A verwandt, aber für ein Alignment zwischen B und C erhalte ich keinen guten Alignment-Score-Wert. Was könnte die Ursache dafür sein? Zeichnen Sie ein Diagramm, falls Ihnen das die Erklärung erleichtert.

Frage 17

Ich habe ein multiples Sequenz-Alignment berechnet. Ich möchte herausfinden, welche Positionen im Alignment konserviert sind und welche variieren. Deshalb erstelle ich eine Abbildung, in welcher die Konservierung bzw. die Variabilität gegen die Sequenzposition aufgetragen wird.

Sie erinnern Sich sicher an die Formel $S = \sum_{i=1}^{N_{states}} p_i \ln p_i$.

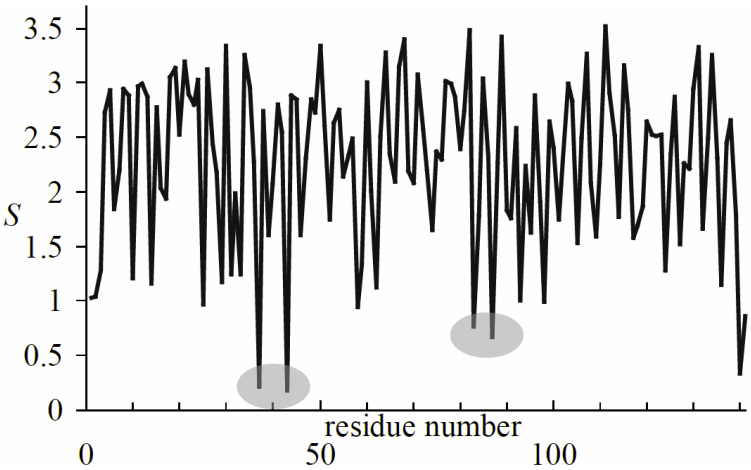
Welche Bedeutung hat p_i in dieser Formel?

Frage 18

Wie würden Sie vorgehen, um in einem multiplen Sequenz-Alignment potentiell katalytisch wichtige Seitenketten im aktiven Zentrum zu identifizieren?
Warum könnten Sie enttäuscht werden?

Frage 19:

Anhand eines multiplen Sequenz-Alignments habe ich Variabilität/Konservierung als eine Funktion der Sequenzposition berechnet.



a) Die Aminosäure-Reste 37 und 43 scheinen sehr gut konserviert zu sein.
Warum könnten diese beiden Reste wichtig sein?

b) Einige Reste in der Abbildung sind nicht besonders gut konserviert. Ich behaupte deshalb, dass diese Reste für die Funktion des Proteins unbedeutend sind.
Warum könnte ich damit falsch liegen?

Frage 20:

Ich habe ein Sequenzalignment von 400 Tyrosin-Kinasen berechnet und stelle fest, dass nur wenige Sequenzpositionen im Verlauf der Evolution konserviert wurden. Was könnte ich an meiner Auswertung ändern, um mehr konservierte Reste zu beobachten?

Frage 21:

Wir haben eine Familie von Sequenzen sowie alle zugehörigen paarweisen Alignments. Für zwei Sequenzen dieser Familie können wir die Anzahl von Unterschieden (Mutationen) zählen und dann mit folgender Formel den Anteil an mutierten Aminosäure-Resten berechnen:

$$p_{mut} = \frac{N_{diff}}{N_{length}}$$

Warum ist dies keine gute Methode?

Frage 22:

Ich möchte überlagerte (alinierte) DNA-Sequenzen für die Erstellung eines phylogenetischen Baumes verwenden. Nennen Sie zwei Gründe dafür, dass die Verzweigungen des Baumes unter Umständen nicht verlässlich sind.

Frage 23:

Ich habe einen phylogenetischen Baum mit einer Neighbor-Joining-Methode erstellt. Wie könnte ich vorgehen, um die Verlässlichkeit des berechneten Baumes zu überprüfen?

Frage 24:

Man nimmt an, dass sich Protein-Sequenzen im Verlauf der Evolution schneller ändern als Proteinstrukturen. Was könnte eine evolutionäre Ursache dafür sein?

Frage 25:

Sie haben ein bakterielles Protein mit unbekannter Funktion. Sie haben eine Knockout-Mutante hergestellt, aber ohne das entsprechende Gen stirbt das Bakterium sofort. Wie würden Sie vorgehen, um etwas über die Funktion dieses Proteins herauszufinden? Nach welcher Art von Information würden Sie suchen?

Frage 26

Warum ist die PDB keine gute Stichprobe für natürliche Proteine?

Frage 27

Welche Darstellungsform würden Sie in *Chimera* für ein Protein auswählen, wenn Sie die Sekundärstruktur sehen möchten?

Frage 28

Die Standardabweichung ist ein Maß für die Streuung einer Wertemenge $(x_1, x_2, ..., x_N)$ um deren Mittelwert $\overline{x}$. Mit der folgenden Formel lässt sich die Standardabweichung s berechnen:

$$s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

Der RMSD-Wert (Root-mean-square deviation) als Maß für den Unterschied zwischen zwei Molekülstrukturen lässt sich mit einer sehr ähnlichen Formel berechnen. Was würde in dieser Formel anstelle von $x_i - \overline{x}$ auftauchen?

Frage 29

Warum ist es schwierig einen RMSD-Wert zu berechnen, wenn zwei Strukturen unterschiedlich groß sind?

Andrew Torda
Björn Hansen
Iryna Bondarenko

Zentrum für Bioinformatik

Übung zur Vorlesung „Bioinformatik“

Sommersemester 2015

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

18.05.2015

Übung 4: Revision

Beispielfragen zur Klausur im Modul „Bioinformatik“ (erste Semesterhälfte)

Anmerkungen:

- Die Prüfungsfragen der Klausur werden auf Deutsch gestellt.
- Sie basieren sowohl auf der Vorlesung als auch auf der Übung.
- Auf den folgenden Seiten finden Sie typische Fragen, wie sie in der Klausur gestellt werden könnten. Dies ist aber kein Fragenkatalog, sondern nur eine kleine Sammlung von möglichen Prüfungsfragen.

Frage 1:

Sie haben zwei DNA-Sequenzen, für die ein Alignment angefertigt werden soll:
ACGTCCTTCATT und GTCTCATG

a) Sie haben das folgende Bewertungsschema:

- Übereinstimmung (match): +1
- Mismatch: -1
- Kosten für das Öffnen einer Lücke (gap opening costs): -1
- Kosten für das Erweitern einer Lücke (gap extension costs): -1

Scheiben Sie das beste lokale Alignment dieser beiden Sequenzen auf.

b) Sie haben das folgende Bewertungsschema:

- Übereinstimmung (match): +1
- Mismatch: 0
- Kosten für das Öffnen einer Lücke (gap opening costs): -10
- Kosten für das Erweitern einer Lücke (gap extension costs): 0

Scheiben Sie das beste lokale Alignment dieser beiden Sequenzen auf.

Frage 2:

Sie haben eine Score-Matrix für ein Paar von DNA-Sequenzen berechnet.
Nach der Traceback-Berechnung erhalten Sie folgendes Ergebnis:

	A	C	A	C	C	T	T	A
C								
C								
A								
T								
C								
C								
A								
A								

Schreiben Sie das zugehörige Sequenzalignment inklusive aller Lücken (gaps) an den richtigen Positionen auf.

Frage 3:

Bei Protein-Alignments werden nicht nur Matches und Missmatches betrachtet, sondern es wird auch die Ähnlichkeit von Aminosäuren berücksichtigt. Wie werden diese Ähnlichkeiten repräsentiert?

Frage 4:

Sie berechnen Alignments von Proteinsequenzen unter Verwendung der folgenden Substitutionsmatrix:

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

a) Kosten für das Öffnen einer Lücke (gap opening costs): -8

Kosten für das Erweitern einer Lücke (gap widening/extension costs): -1

Sie erhalten das folgende Alignment:

AACDQRST
A-CD-RST

Wie lautet der Score-Wert dieses Alignments?

b) Angenommen Sie bekämen das folgende Alignment:

AACDQRST
A-CD--ST

Wie lautete jetzt der Score-Wert?

c)

AACDQRST
A-CD-SST

Wie lautet der Score-Wert für dieses Alignment?

Frage 5:

Für Berechnung von Proteinsequenz-Alignments können viele verschiedene Substitutionsmatrizen verwendet werden. Warum könnte eine bestimmte Substitutionsmatrix gegenüber einer anderen bevorzugt werden?

Frage 6:

Oft kommt die Substitutionsmatrix BLOSUM zum Einsatz.
Skizzieren Sie die nötigen Schritte zur Berechnung der Werte in dieser Matrix.

Frage 7:

a) Worin besteht der Vorteil eines Needleman-Wunsch-Alignments gegenüber einem „seeded"-Alignment" (z.B. einem BLAST-Alignment)?

b) Was ist der Vorteil einer „seeded"-Methode" wie BLAST gegenüber dem Needleman-Wunsch-Verfahren?

c) Nennen Sie eine Anwendung, bei welcher Sie Alignments bevorzugt mit BLAST und nicht mit dem Needleman-Wunsch-Verfahren berechnen würden.

d) Nennen Sie eine Anwendung, bei welcher Sie das eher langsame Needleman-Wunsch-Verfahren einem BLAST-ähnlichen Verfahren vorziehen würden.

Frage 8:

a) Was ist der Unterschied zwischen einer wiederholt durchgeführten BLAST-Suche (PSI-BLAST) und einer gewöhnlichen BLAST-Suche?

b) Was ist der Vorteil von PSI-BLAST gegenüber einer gewöhnlichen BLAST-Suche?

Frage 9:

Sie haben zwei entfernt verwandte Proteine mit schwacher Sequenzähnlichkeit. Sie kennen sowohl die zugehörigen Protein- als auch die DNA-Sequenzen. Wieso würden Sie ein besseres Alignment erwarten, wenn Sie die Proteinsequenzen für das Alignment verwenden?

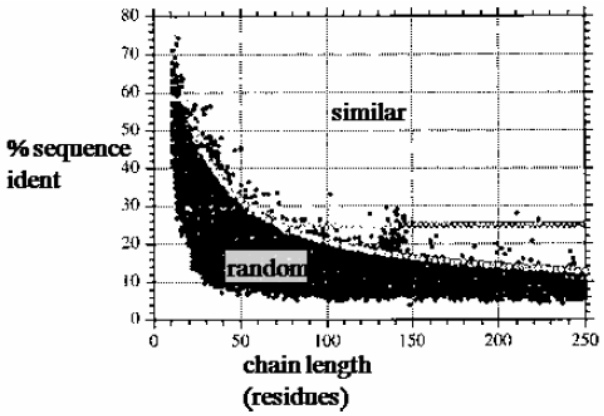
Frage 10:

a) Ich habe ein Programm erstellt, welches DNA-Zufallssequenzen erzeugt. Für paarweise Alignments solcher Sequenzen erwarte ich eine durchschnittliche Sequenzübereinstimmung von ca. 25%. Bei der Berechnung stelle ich aber fest, dass die Sequenzübereinstimmung meistens höher als 25% ist. Was könnte der Grund dafür sein?

b) Wenn die Sequenzen nicht zufällig erzeugt, sondern stattdessen zufällig aus einer Datenbank für biologische Sequenzen ausgewählt werden, ist die beobachtete Sequenzähnlichkeit sogar noch größer. Wie ist das zu erklären?

Frage 11:

a) Wenn wir für alle strukturell unterschiedlichen Paare von Proteinen aus einer großen Menge von Proteinen (z.B. der PDB) die Kettenlänge gegen die Sequenzähnlichkeit auftragen, dann entsteht ein Diagramm wie das Folgende:



Welche Aussagen lassen sich anhand dieser Abbildung treffen?

b) Ich habe zwei Proteine mit einer Sequenzübereinstimmung von 20% und möchte wissen, wie signifikant diese Sequenzähnlichkeit ist. Welche andere Information benötigte ich noch, um diese Frage beantworten zu können?

Frage 12

a) Was zeichnet ein korrektes Alignment aus?
Geben Sie eine biologisch und eine algorithmisch/mathematisch begründete Antwort!

b) Häufig ist das Finden des korrekten Alignments erschwert. Nennen und erläutern Sie drei Gründe dafür!

Frage 13

Ich möchte ein multiples Sequenzalignment für N_{seq} Sequenzen berechnen. Wie viele paarweise Alignments muss ich hierzu berechnen?

Frage 14

In einem multiplen Sequenz-Alignment wollen sie einen Score-Wert maximieren:

$$score = \sum_{b \neq \grave{a}}^{N_{seq}} \sum_{a=1}^{N_{seq}} S_{a,b}$$

Was bedeutet dieser Score-Wert in Worten?

Frage 15

Ich möchte einen Leitbaum (guide tree) für ein multiples Sequenzalignment erstellen. Wodurch wird die Reihenfolge festgelegt, in welcher die Knoten des Baumes verknüpft werden?

Frage 16

Ich habe 3 Sequenzen: A, B und C. Die Sequenzen B und C sind beide mit A verwandt, aber für ein Alignment zwischen B und C erhalte ich keinen guten Alignment-Score-Wert. Was könnte die Ursache dafür sein? Zeichnen Sie ein Diagramm, falls Ihnen das die Erklärung erleichtert.

Frage 17

Ich habe ein multiples Sequenz-Alignment berechnet. Ich möchte herausfinden, welche Positionen im Alignment konserviert sind und welche variieren. Deshalb erstelle ich eine Abbildung, in welcher die Konservierung bzw. die Variabilität gegen die Sequenzposition aufgetragen wird.

Sie erinnern Sich sicher an die Formel $S = \sum_{i=1}^{N_{states}} p_i \ln p_i$.

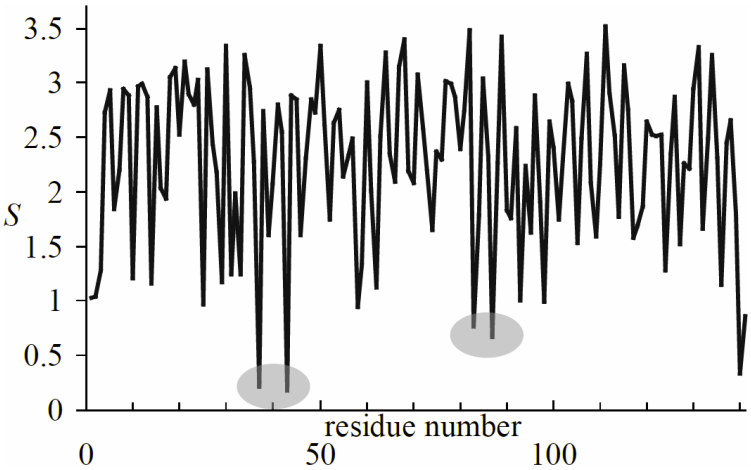
Welche Bedeutung hat p_i in dieser Formel?

Frage 18

Wie würden Sie vorgehen, um in einem multiplen Sequenz-Alignment potentiell katalytisch wichtige Seitenketten im aktiven Zentrum zu identifizieren?
Warum könnten Sie enttäuscht werden?

Frage 19:

Anhand eines multiplen Sequenz-Alignments habe ich Variabilität/Konservierung als eine Funktion der Sequenzposition berechnet.



a) Die Aminosäure-Reste 37 und 43 scheinen sehr gut konserviert zu sein.
Warum könnten diese beiden Reste wichtig sein?

b) Einige Reste in der Abbildung sind nicht besonders gut konserviert. Ich behaupte deshalb, dass diese Reste für die Funktion des Proteins unbedeutend sind.
Warum könnte ich damit falsch liegen?

Frage 20:

Ich habe ein Sequenzalignment von 400 Tyrosin-Kinasen berechnet und stelle fest, dass nur wenige Sequenzpositionen im Verlauf der Evolution konserviert wurden. Was könnte ich an meiner Auswertung ändern, um mehr konservierte Reste zu beobachten?

Frage 21:

Wir haben eine Familie von Sequenzen sowie alle zugehörigen paarweisen Alignments. Für zwei Sequenzen dieser Familie können wir die Anzahl von Unterschieden (Mutationen) zählen und dann mit folgender Formel den Anteil an mutierten Aminosäure-Resten berechnen:

$$p_{mut} = \frac{N_{diff}}{N_{length}}$$

Warum ist dies keine gute Methode?

Frage 22:

Ich möchte überlagerte (alinierte) DNA-Sequenzen für die Erstellung eines phylogenetischen Baumes verwenden. Nennen Sie zwei Gründe dafür, dass die Verzweigungen des Baumes unter Umständen nicht verlässlich sind.

Frage 23:

Ich habe einen phylogenetischen Baum mit einer Neighbor-Joining-Methode erstellt. Wie könnte ich vorgehen, um die Verlässlichkeit des berechneten Baumes zu überprüfen?

Frage 24:

Man nimmt an, dass sich Protein-Sequenzen im Verlauf der Evolution schneller ändern als Proteinstrukturen. Was könnte eine evolutionäre Ursache dafür sein?

Frage 25:

Sie haben ein bakterielles Protein mit unbekannter Funktion. Sie haben eine Knockout-Mutante hergestellt, aber ohne das entsprechende Gen stirbt das Bakterium sofort. Wie würden Sie vorgehen, um etwas über die Funktion dieses Proteins herauszufinden? Nach welcher Art von Information würden Sie suchen?

Frage 26

Warum ist die PDB keine gute Stichprobe für natürliche Proteine?

Frage 27

Welche Darstellungsform würden Sie in *Chimera* für ein Protein auswählen, wenn Sie die Sekundärstruktur sehen möchten?

Frage 28

Die Standardabweichung ist ein Maß für die Streuung einer Wertemenge $(x_1, x_2, ..., x_N)$ um deren Mittelwert $\overline{x}$. Mit der folgenden Formel lässt sich die Standardabweichung s berechnen:

$$s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

Der RMSD-Wert (Root-mean-square deviation) als Maß für den Unterschied zwischen zwei Molekülstrukturen lässt sich mit einer sehr ähnlichen Formel berechnen. Was würde in dieser Formel anstelle von $x_i - \overline{x}$ auftauchen?

Frage 29

Warum ist es schwierig einen RMSD-Wert zu berechnen, wenn zwei Strukturen unterschiedlich groß sind?