Workshop: Eine Entdeckungsreise in die Welt der Gene

Abteilung für Genominformatik
Zentrum für Bioinformatik
Universität Hamburg
Bundesstrasse 43
20146 Hamburg
http://www.zbh.uni-hamburg.de
→Studium→Informationen für Schüler/-innen→ Workshop DNA-Sequenzanalyse

20. April 2017

Dieser Workshop wurde im Februar 2009 von Ute Willhoeft und Sascha Steinbiss im Rahmen des Schülerkongresses „Evolution Heute, Hamburger Wissenschaftstage“ zum Darwinjahr 2009 entwickelt. Ein Überarbeitung und Neugestaltung des Dokumentes erfolgte im März/April 2017 durch Stefan Kurtz und Annika Seidel. Die Screenshots wurden unter MS-Windows 8 erstellt. Sie weichen daher möglicherweise leicht von der Darstellung unter Linux ab.

Vorwort

Die Entdeckungsreise ist für Schüler (Jahrgang 11 und 12) und interessierte Laien so verfasst, dass die Reise auch zu Hause durchgeführt werden kann.

Dieser Workshop gibt Beispiele, wie Bioinformatik-Software von Wissenschaftlern im Bereich Life Science (Biologie, Biochemie, Medizin) genutzt wird. Hier ist nur ein kleiner Ausschnitt aus der anwendungs-orientierten Bioinformatik gezeigt. Bioinformatik als Fach umfasst natürlich weitere Gebiete. Weitere Informationen zu Forschungsinhalten der Bioinformatik und zu Studienmöglichkeiten an der Universität Hamburg sind unter `http://zbh.uni-hamburg.de` zu finden.

Wenn Sie dieses Dokument am Rechner ansehen, können Sie mit einem geeigneten Reader (z.B. dem acrobat reader) die Hyperlinks verfolgen.

Dieses Dokument enthält mehrere Abschnitte:

- Der Reiseführer (Abschnitt 2) liefert die Hintergrundinformationen. Hier kann z.B. nachgelesen werden, wie Gene aufgebaut sind, was eine Alignment ist und wie man die benötigten Programme installiert. Der Reiseführer sollte nur dann gelesen werden, wenn etwas unklar ist oder wenn etwas so spannend ist, dass man gern Zusatzinformationen hätte.
- Die Landkarte gibt es mit zwei Touren, nämlich
  - 1.1→1.2→1.3 und
  - 1.1→1.2→1.4→1.5.

  Sie ist die Wegbeschreibung und verweist an vielen Stellen auf den Reiseführer, der dann die Sehenswürdigkeiten erläutert.

Für die Entdeckungsreise benötigt man:

- Neugierde,
- einen Computer mit Internetzugang,
- die Programme ClustalX und Dendroscope, deren Installation in Abschnitt 2.16 bzw. 2.17 beschrieben wird[1],
- Englischkenntnisse (Mittelstufenniveau),
- 30 Minuten Zeit für das Stöbern im Reiseführer,
- 1-2 Stunden Zeit für die Aufgaben der Tour 1.1→1.2→1.3
- 1-2 Stunden Zeit für die Aufgaben der Tour 1.1→1.2→1.4→1.5, wobei 1.5 auch weggelassen werden kann.

Viel Spaß wünscht die Abteilung für Genominformatik am Zentrum für Bioinformatik der Universität Hamburg.

Danksagung

Vielen Dank an Malik Alawi und Joachim Trucks für die hilfreichen Anmerkungen und Diskussionen bei der Erstellung dieses Workshops, an Adrian Kolodzik für die Idee, *Myostatin* als Beispiel zu verwenden und an Karin Lundt für Hinweise zu sprachlichen und inhaltlichen Verbesserungen.

[1] Die Installation der Programme ist nicht notwendig, wenn Sie den Workshop an den Rechnern im Zentrum für Bioinformatik durchführen.

Inhaltsverzeichnis

1 Landkarte für Pauschalreisende

Hinweise:

- Zunächst sollten die Programme ClustalX und Dendroscope installiert werden, siehe Abschnitt 2.16 bzw. 2.17.[2]
- In der Landkarte findet man eine detaillierte Anleitung zur Vorgehensweise und Fragen zu den einzelnen Punkten. Hintergrundwissen kann jeweils nachgelesen werden, wenn man den Verweisen auf die entsprechenden Abschnitt im Reiseführer folgt. Beispiel: Die Informationen zum Gen *Myostatin*, dem Fallbeispiel in diesem Workshop, findet man im Abschnitt 2.12.
- Die Tour 1.1 führt in die Programme ClustalX und Dendroscope ein und zeigt, wie das Programm ein Multiples Sequenzalignment (MSA) erstellt und wie man das Programm benutzt. Hier sind die einzelnen Schritte detailliert beschrieben, während die weiteren Touren dann nicht mehr detailliert auf die Programmnutzung eingehen, sondern im Wesentlichen die Fragen aufzeigen, die man mit Sequenzalignments klären kann und z.T. auch Antworten beinhalten.
- Zwei Touren stehen zur Auswahl:
  - Auf Tour 1.1→1.2→1.3 ist man einem Gen von der DNA über die mRNA bis zum Protein auf der Spur.
  - Auf Tour 1.1→1.2→1.4→1.5 werden Stammbäume auf der Basis von Proteinsequenzen erstellt und Proteinfamilien entdeckt.
- Für die einzelnen Abschnitte der Tour ist jeweils neben der Überschrift die ungefähre Dauer in Minuten angegeben.

Für die Entdeckungsreise werden die die folgenden Dateien benötigt:

- `mstn-mRNA.fas`
- `mstn-protein.fas`
- `mstn-genome.fas`
- `unknown-proteins.fas`
- `bmp2-protein.fas`
- `mstn-bmp2-protein.fas`
- `Fruchtfliege-protein.fas`
- `Gelbfiebermuecke-protein.fas`

Diese Dateien stehen unter `http://www.zbh.uni-hamburg.de` im Bereich

Studium→Informationen für Schüler/-innen→ Workshop DNA-Sequenzanalyse

zum Download als Zip-Archiv zur Verfügung. Dieses kann man mit Standardprogrammen einfach entpacken. Eine kurze Einführung zu DNA-Sequenzen und deren Speicherung in Dateien ist hilfreich, um die Entdeckungsreise entspannt zu beginnen.

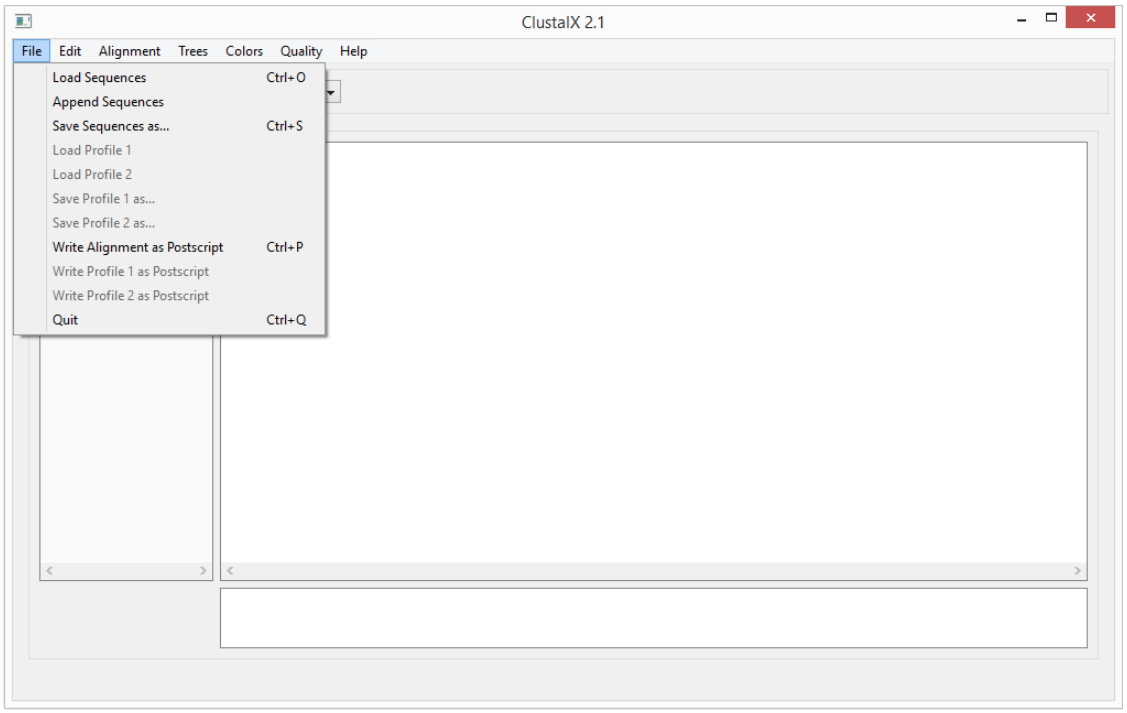
[2] Die Installation die Programme ist nicht notwendig, wenn Sie den Workshop an den Rechnern im Zentrum für Bioinformatik durchführen.

1.1 Mit ClustalX ein Multiples Alignment erstellen

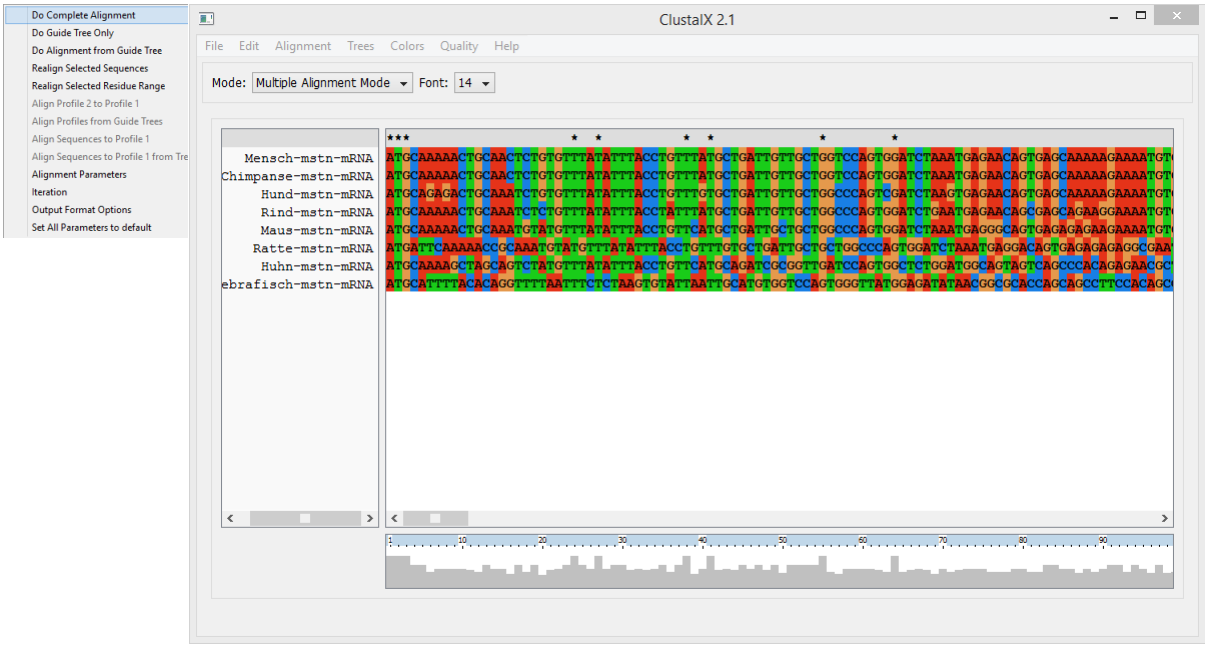
20-30 Min.

Als erste Aufgabe soll mit dem Programm ClustalX ein Multiples Sequenzalignment (MSA) von mRNA-Sequenzen des Gens *Myostatin* aus verschiedenen Arten erstellt werden.

Man startet zunächst ClustalX und es erscheint die folgende Ansicht:

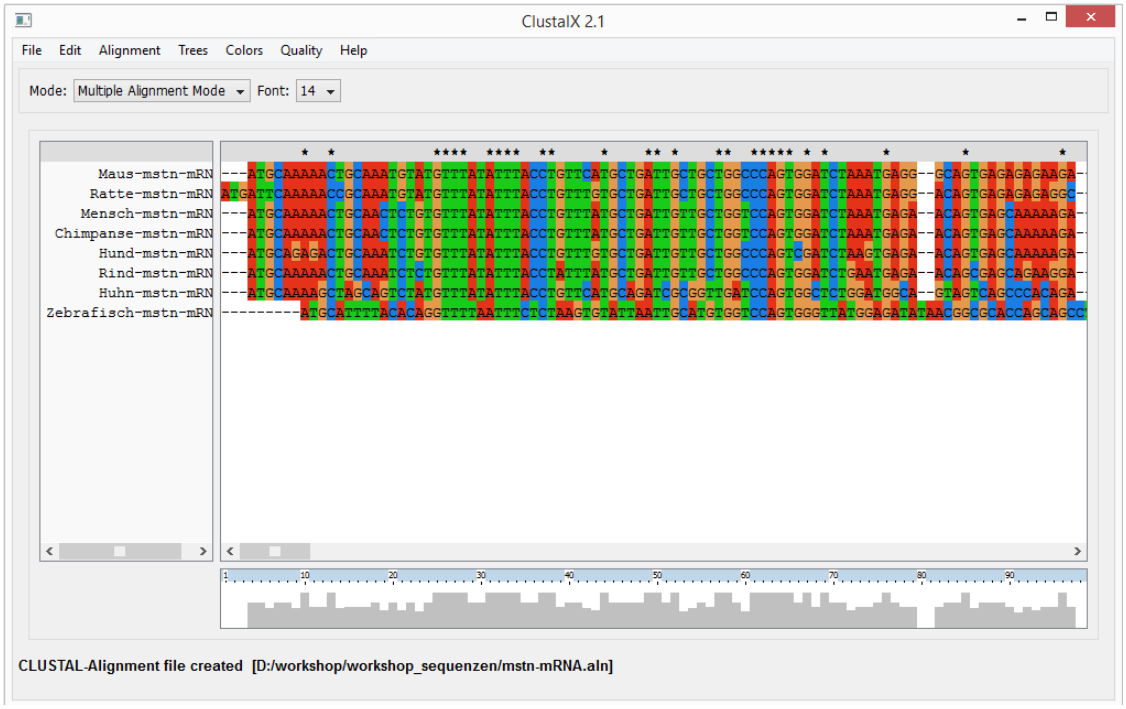


Über das Menü *File* können Dateien geladen und gespeichert werden. Auf *Load sequences* klicken und die Datei `mstn_mRNA.fas` auswählen. Es erscheint:



Über das Menü *Alignment* werden die geladenen Sequenzen miteinander verglichen. Nach Klick auf *Do Complete Alignment* dauert es eine kurze Zeit bis das Alignment erscheint. In der Konsole steht dann `CLUSTAL-Alignment file created [...]`. Über die Choice-Box *Font* kann die Grösse der Buchstaben im Alignment verändert werden.

Das berechnete und in `mstn-mRNA.aln` gespeicherte Alignment sieht so aus:



Andere Menüpunkte werden später noch eine Rolle spielen. Sollten die vier Basen nicht in vier verschiedenen Farben angezeigt werden, dann das Menü *Color* und *Load Color Parameter File* klicken. Hier die Datei `coldna.xml` bzw. für Proteine `colprot.xml` laden. Diese Dateien befinden sich in dem Ordner, in dem ClustalX installiert ist.

1.1.1 Wie sieht ein Alignment von mRNA-Sequenzen aus?

Hintergrundwissen findet man in Abschnitt 2.11.

Versuchen Sie Antworten auf die folgenden Fragen zu finden:

1. Welche Bereiche im Alignment passen gut zueinander?
2. Wo unterscheiden sich die Sequenzen voneinander?
3. Welche Sequenz unterscheidet sich deutlich von den anderen Sequenzen?

Die $*$-Symbole in der oberen Zeile und grauen Balken unten helfen das Alignment zu bewerten.

1.1.2 Wie hat das Programm ClustalX das Alignment erstellt?

Versuchen Sie, zunächst selbst zu verstehen, was das Programm gemacht hat. Hintergrundwissen findet man in Abschnitt 2.11. Während der Berechnung werden von ClustalX die wesentlichen Schritte durch Ausgabe in der Konsole dokumentiert. Das sieht dann so aus.

```
Sequences assumed to be DNA
Start of Pairwise alignments
Aligning...

Sequences (1:2) Aligned. Score:  99
Sequences (1:3) Aligned. Score:  91
Sequences (1:4) Aligned. Score:  91
Sequences (1:5) Aligned. Score:  91
Sequences (1:6) Aligned. Score:  90
Sequences (1:7) Aligned. Score:  85
Sequences (1:8) Aligned. Score:  65
Sequences (2:3) Aligned. Score:  91
Sequences (2:4) Aligned. Score:  91
Sequences (2:5) Aligned. Score:  92
Sequences (2:6) Aligned. Score:  90
Sequences (2:7) Aligned. Score:  84
Sequences (2:8) Aligned. Score:  65
Sequences (3:4) Aligned. Score:  91
Sequences (3:5) Aligned. Score:  89
Sequences (3:6) Aligned. Score:  89
Sequences (3:7) Aligned. Score:  82
Sequences (3:8) Aligned. Score:  64
Sequences (4:5) Aligned. Score:  89
Sequences (4:6) Aligned. Score:  88
```

```
Sequences (4:7) Aligned. Score:  81
Sequences (4:8) Aligned. Score:  65
Sequences (5:6) Aligned. Score:  96
Sequences (5:7) Aligned. Score:  83
Sequences (5:8) Aligned. Score:  60
Sequences (6:7) Aligned. Score:  82
Sequences (6:8) Aligned. Score:  65
Sequences (7:8) Aligned. Score:  61
Guide tree file created:   [mstn-mRNA.dnd]

There are 7 groups
Start of Multiple Alignment

Aligning...
Group 1: Sequences:   2      Score:21403
Group 2: Sequences:   2      Score:20273
Group 3: Sequences:   4      Score:20322
Group 4: Sequences:   2      Score:20919
Group 5: Sequences:   6      Score:20026
Group 6: Sequences:   7      Score:19008
Group 7: Sequences:   8      Score:15446
Alignment Score 176316

CLUSTAL-Alignment file created  [mstn-mRNA.aln]
```

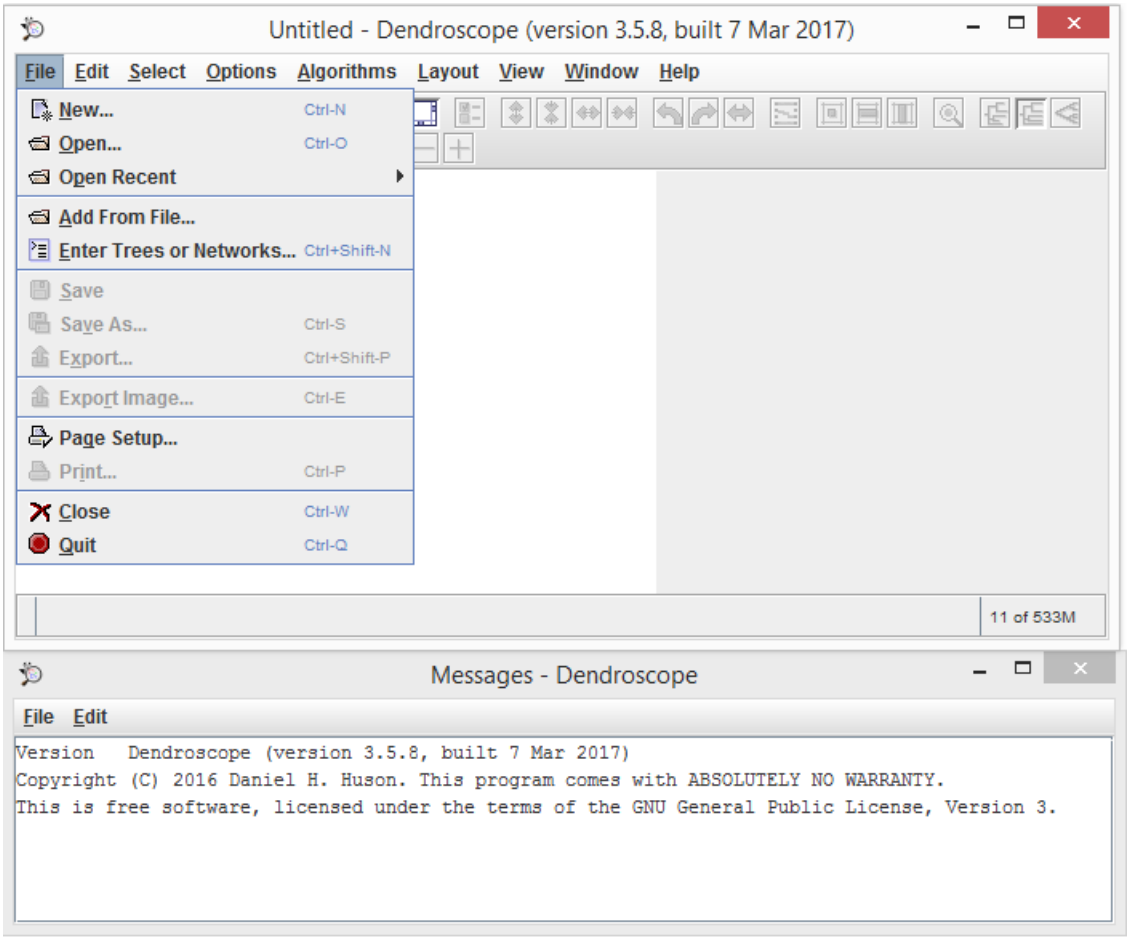
1.1.3 Welche Informationen sind im Leitbaum vorhanden?

Der Leitbaum ist ein Stammbaum, der vom Programm errechnet wurde. Die Dateiendung ist `.dnd`, also z.B. `mstn-mRNA.dnd` für die Berechnung unter 1.1.1.

Um sich diesen Baum anzusehen, startet man das Programm Dendroscope. Auch dieses Programm ist (fast) selbsterklärend. Benötigt wird das Menü *File* zum Laden und Speichern von Dateien sowie *Layout* für die Auswahl der Darstellungsformen Phylogramm und Cladogramm. Dabei sollte darauf geachtet werden, dass unter *Files of Type* die die Option *all files* aktiviert ist.

Das Menü *Tree* bietet verschiedene Ansichten eines Baums. Die Ansicht *Rectangular Phylogram* ist für diesen Workshop sinnvoll. Sollte nicht der gesamte Baum angezeigt sein, dann unter *View* auf *Expand tree* klicken.

Was ist der Unterschied zwischen Phylogramm und Cladogramm?



Tipp: Unter dem Menü *Layout* beide Ansichten anzeigen lassen und miteinander vergleichen.

Welche Informationen sind im Leitbaum vorhanden und wie werden diese vom Programm Dendroscope genutzt, um den Stammbaum zu zeichnen? Als Beispiel ist hier der Leitbaum aus Aufgabe 1.1.1 dargestellt:

```
(
(
(
Mensch-mstn-mRNA:0.00124,
Chimpanse-mstn-mRNA:0.00053)
:0.03003,
(
Hund-mstn-mRNA:0.04226,
Rind-mstn-mRNA:0.04373)
:0.00743)
:0.00209,
(
```

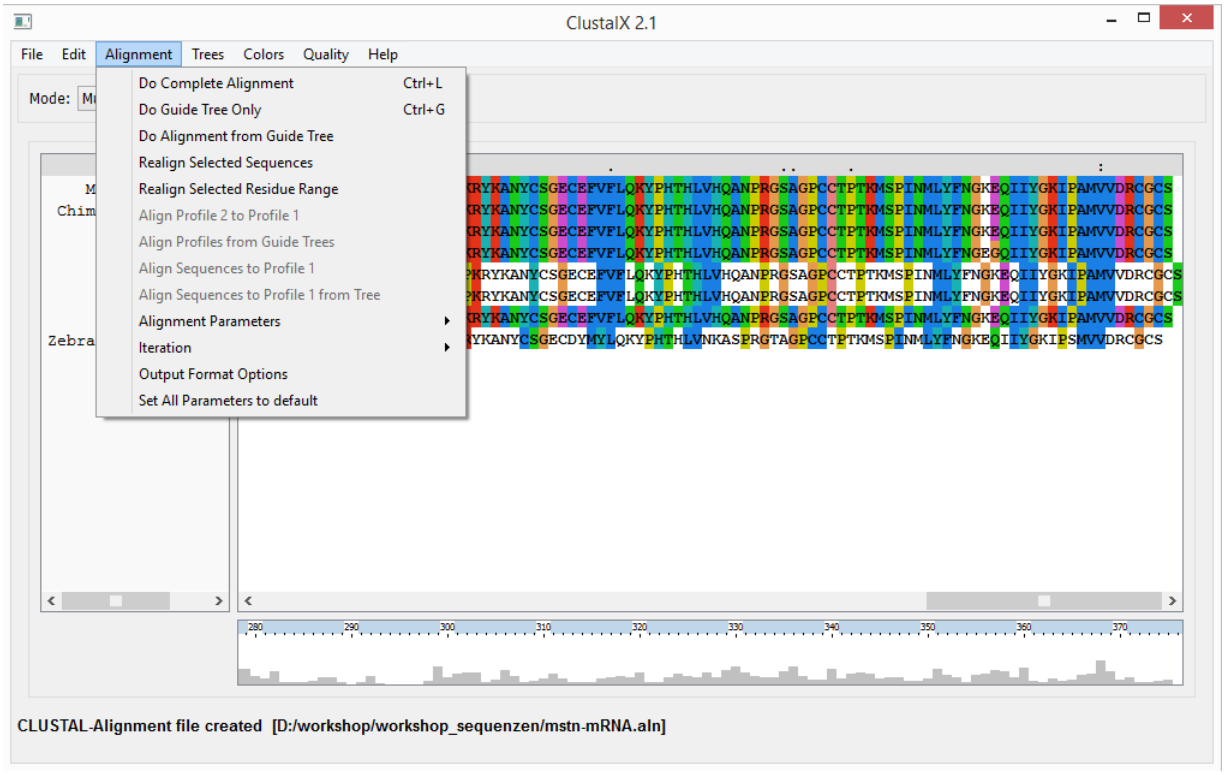
```
Maus-mstn-mRNA:0.01804,
Ratte-mstn-mRNA:0.01742)
:0.03736,
(
Huhn-mstn-mRNA:0.09911,
Zebrafisch-mstn-mRNA:0.28311)
:0.02224);
```

Tipp: In einem Editor (z.B. KWrite unter Linux oder WordPad unter Windows) die Datei `mstn-mRNA.dnd` öffnen und hineinsehen.

1.2 Erstellen eines MSA der Proteinsequenzen von *Myostatin*

10-15 Min.

1.2.1 Für die Sequenzen aus `mstn-protein.fas` ein MSA erstellen

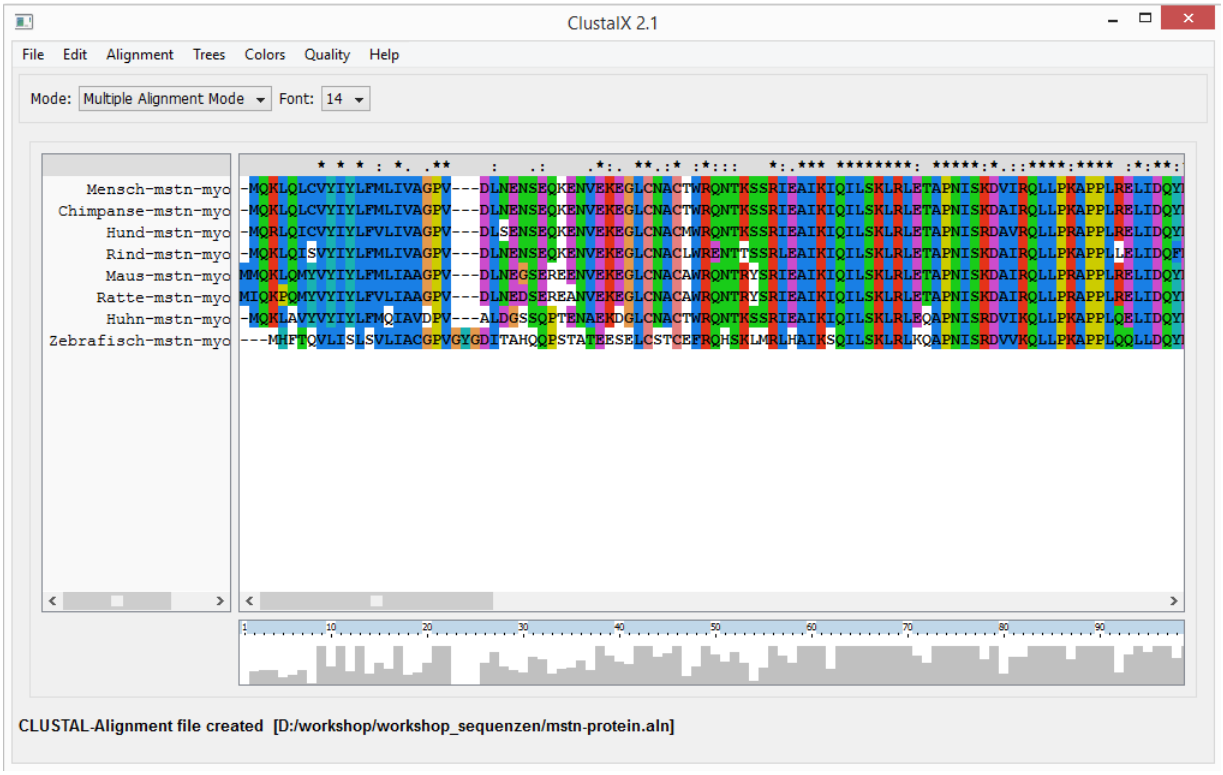


Sollten die Aminosäuren farblich nicht wie in Abbildungen angezeigt sein, dann das Menü *Color* und *Load Color Parameter File* wählen. Hier die Datei `colprot.xml` laden. Sie befindet sich in dem Ordner, in dem ClustalX installiert ist.

Die Darstellung des Proteinsequenz-Alignments ist genauso wie die Darstellung des DNA-Sequenz-Alignments. Statt 4 Basen stehen natürlich die 1-Buchstaben Codes der 20 Ami-

nosäuren in den Alignmentspalten. Über dem Alignment sind Spalten mit den Symbolen $*$, **:** oder **.** gekennzeichnet. Was könnten diese Zeichen bedeuten?

Tipp: Unter *Help* →*General* findet man die Erklärung.



1.2.2 Vergleich der Alignments

Vergleichen Sie nun die MSAs der mRNA-Sequenzen und der Proteinsequenzen. In welchem der beiden Alignments sieht man (relativ zur Länge) mehr konservierte Spalten, d.h. Spalten mit identischen Symbolen? Sie müssen hier keine genaue Zählung vorzunehmen, sondern sich für eine der beiden Möglichkeiten entscheiden und dafür einen guten Grund angeben.

1.3 Erstellen eines MSA der genomischen Sequenz von *Myostatin* und Vergleich mit der mRNA

45-60 Min.

1.3.1 Erstellen eines MSA für die Sequenzen aus `mstn-genome.fas`

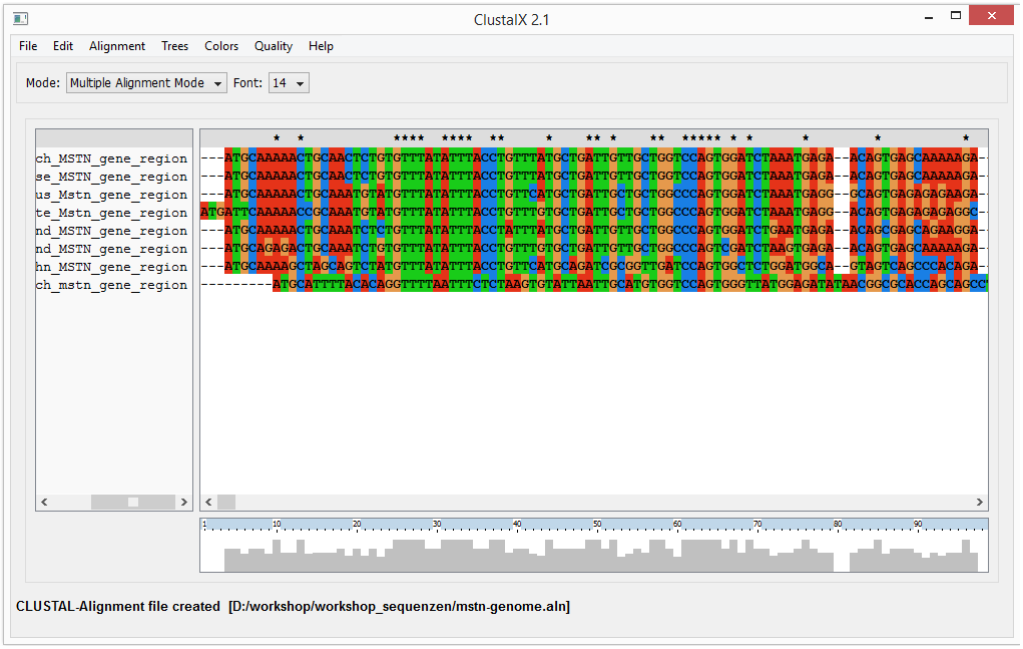
Siehe hierzu auch Abschnitt 1.1.1. Allerdings wird hier ein MSA der genomischen Sequenzen berechnet, d.h. Sie laden die Datei `mstn-genome.fas`. Da die genomischen Sequenzen

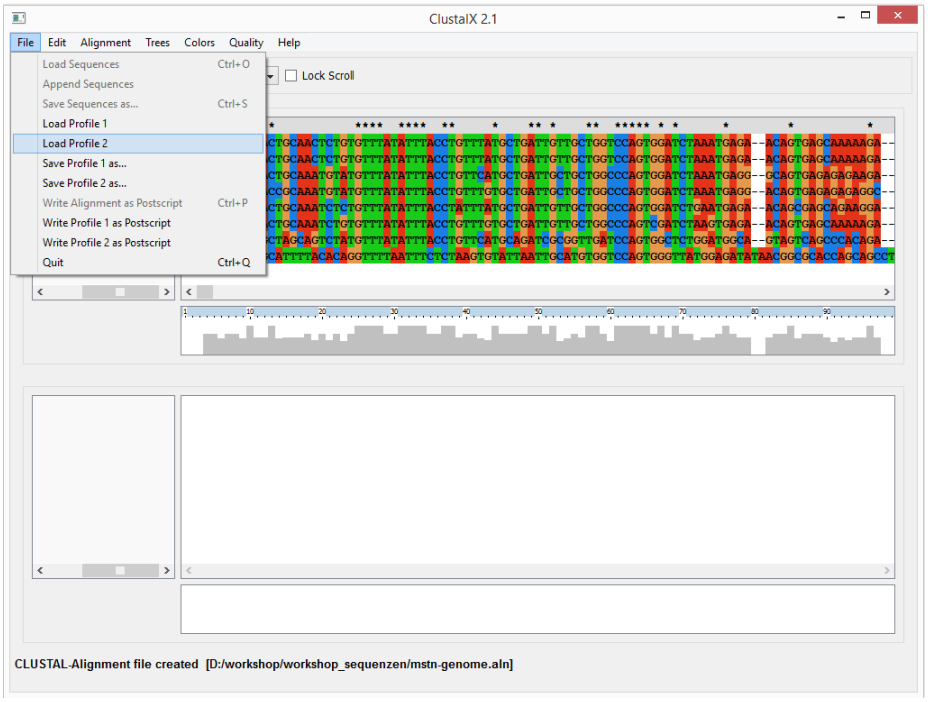
länger sind als die entsprechenden mRNA-Sequenzen, dauert das etwas länger, ca. 1-2 Minuten.

Nach der Berechnung des Alignments (in der Konsole erscheint dann `CLUSTAL-Alignment file created [...]`) wechselt man den *Mode* auf *Profile Alignment Mode*. Man kann nun als zweites Profil das bereits in Abschnitt 1.1.1 berechnete MSA aus Datei `mstn-mRNA.aln` laden.

Tipp: Als zweites Profil die Datei mit der Endung `.aln` verwenden und nicht die mit der Endung `.fas`.

Damit das Alignment so aussieht wie im Screenshot, muss ggf. der Schieberegler unterhalb des Alignmentbereichs verwendet werden.





1.3.2 Vergleich der genomischen DNA und der mRNA

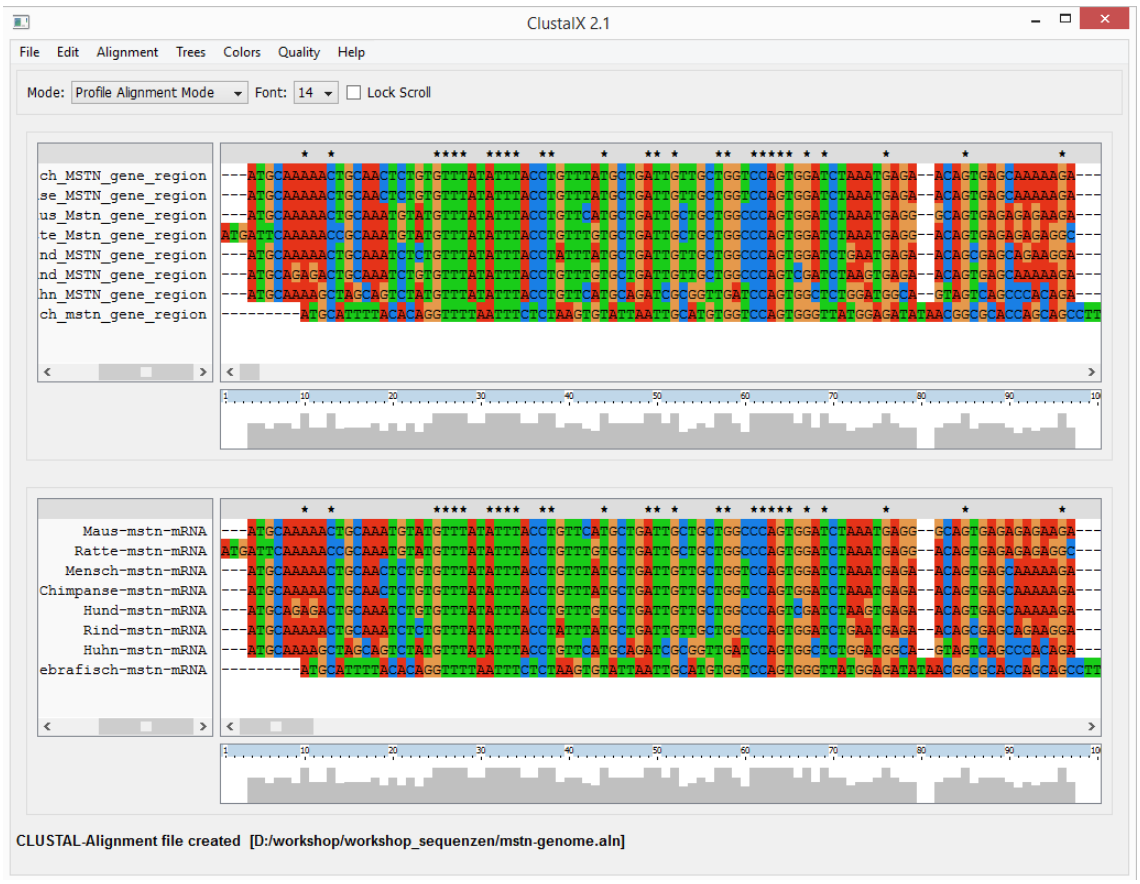
Der Vergleich zeigt, das Sequenzabschnitte aus der DNA in der mRNA fehlen. Wie heißt der liegende Prozess?

Tipp: Auf die Gesamtlänge der Sequenzen und die unterschiedliche Konservierung achten. Hintergrundinformationen sind im Reiseführer im Abschnitt 2.13 zu finden.

1.3.3 Konservierung der Sequenzen

Gibt es Unterschiede im Grad der Konservierung zwischen Sequenzen, die nur in der genomischen DNA bzw. Sequenzen, die sowohl in der DNA als auch der mRNA vorkommen?

Tipp: In diesem Beispiel sieht man sehr schön, was Mutation und Selektion bewirken.



1.3.4 Anzahl der Exons

Bestimmen Sie die Anzahl der Exons des *Myostatin*-Gens.

1.3.5 Exon/Intron Grenzen

Kann man die Grenzen zwischen Exon und Intron im MSA erkennen?

Tipp: Ein Intron beginnt in den meisten Fällen mit `GT` und endet mit `AG`.

1.4 Neue Arten in die Stammbäume einfügen

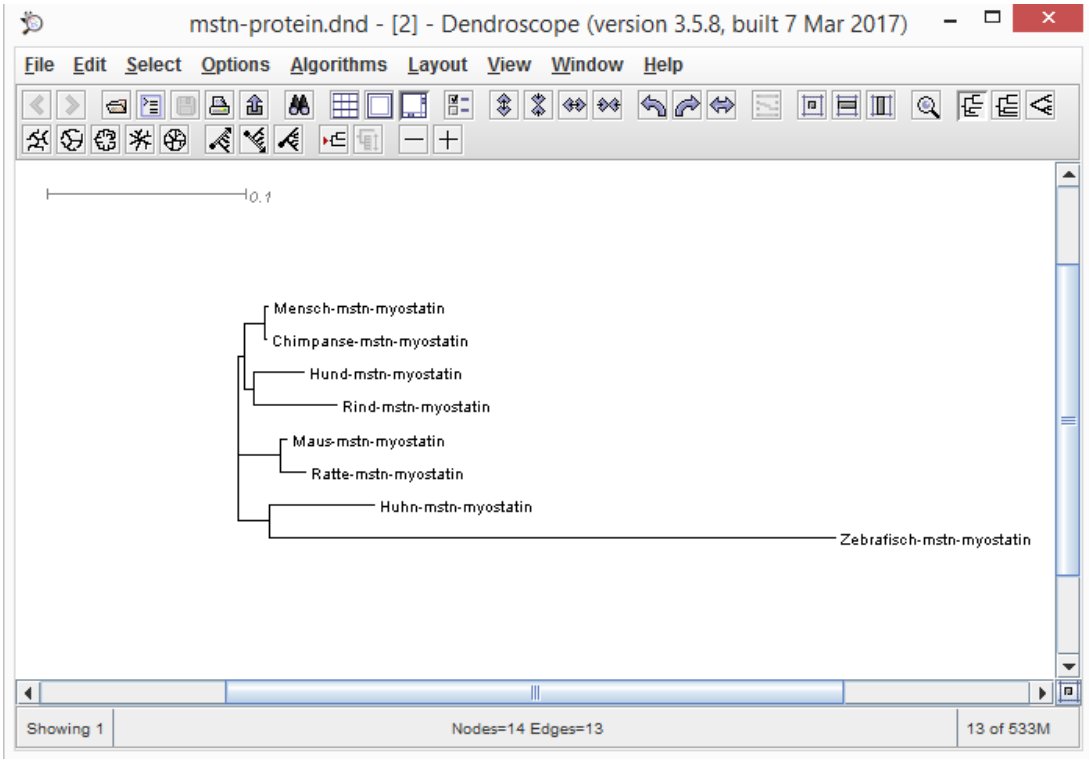
30-45 Min.

In Tour 1.1 wurde bereits gezeigt, dass man mit Multiplen Sequenzalignments (MSA) nicht nur die Unterschiede zwischen Sequenzen offenlegen, sondern auch Stammbäume erstellen kann. Die Stammbäume sind in dem Leitbaum (in der Datei mit der Endung `.dnd`) gespeichert und können mit dem Programm Dendroscope angesehen werden. In diesem Workshop benutzen wir nur die einfachste Methode, um Stammbäume zu erstellen. In der Forschung

sind die Fragestellungen ähnlich zu denen in unserem Workshop. Zur Lösung der Fragen werden aber meist mehrere und komplexere Methoden zur Stammbaumerstellung benutzt. Hier nutzen wir ClustalX, um typische Fragestellungen aus der Forschung zu bearbeiten. Wir haben in Tour 1.1 bereits einen Stammbaum mit den Daten zu einem bestimmten Protein erstellt.

1.4.1 Für die Sequenzen aus `mstn-protein.fas` ein MSA erstellen

Das Alignment ist bereits unter 1.2 bzw. in der Datei `mstn-protein.aln` zu finden. Mit dem Programm Dendroscope den Stammbaum ansehen. Durch Auswahl der Ansicht *Phylogram* werden die Astlängen skaliert, also abhängig von der Anzahl der Unterschiede im Alignment ausgegeben. Das Ergebnis sieht so aus:



Der Stammbaum enthält verschiedene Wirbeltierarten. Entspricht die Anordnung dem, was man erwarten würde?

Sind z.B. die Nagetiere und die Affenartigen Spezies jeweils am engsten miteinander verwandt? Wieso sind die Astlängen zum Zebrafisch und zum Huhn besonders lang?

Ja, hierzu muss man sich ein bisschen in der Systematik der Tiere auskennen. Aber wir haben absichtlich Arten gewählt, die jeder kennt.

1.4.2 Neustart von ClustalX und Öffnen der Datei `mstn-protein.fas`

Über das Menü *File/Append Sequences* die Datei `unknown-proteins.fas` hinzufügen.

Im unteren Bereich sind 5 neue Sequenzen eingefügt.

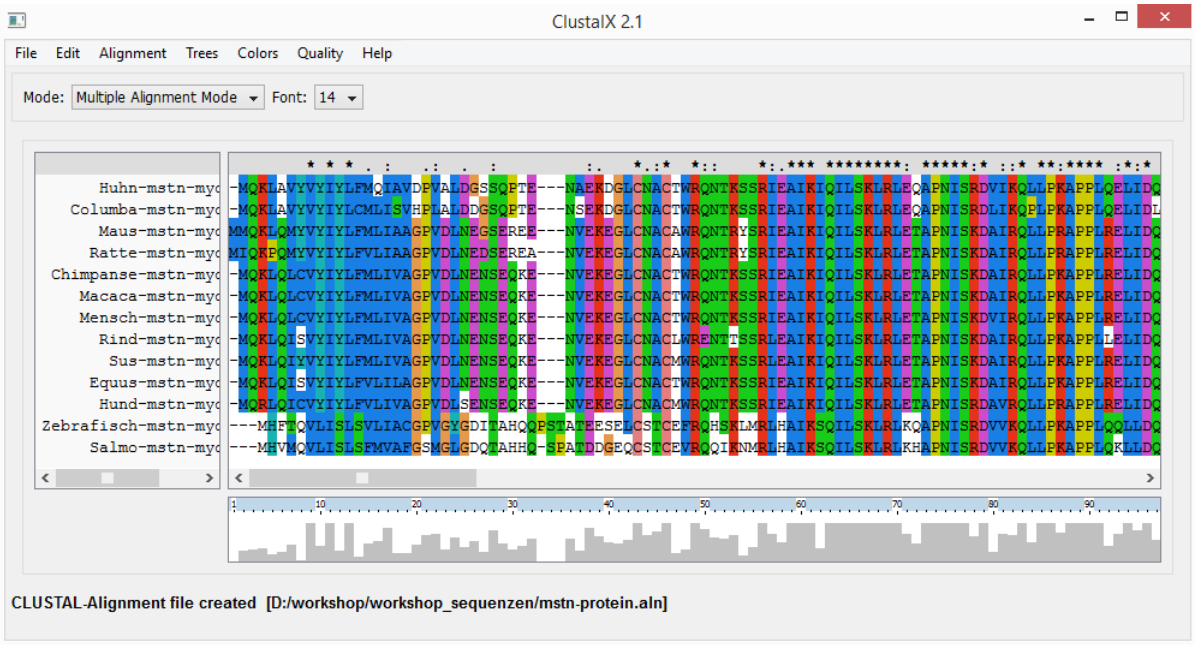
1.4.3 Erstellen von Alignment und Leitbaum

Benutzen Sie Dendroscope, um die Phylogramme darzustellen. Wenn nicht alle, sondern nur einige Arten zum Stammbaum hinzugefügt werden sollen, dann können ggf. Arten durch Klicken auf einzelne Sequenzen unter dem Menü *Edit/Clear sequence selection* entfernt werden, bevor MSA und Leitbaum erstellt werden.

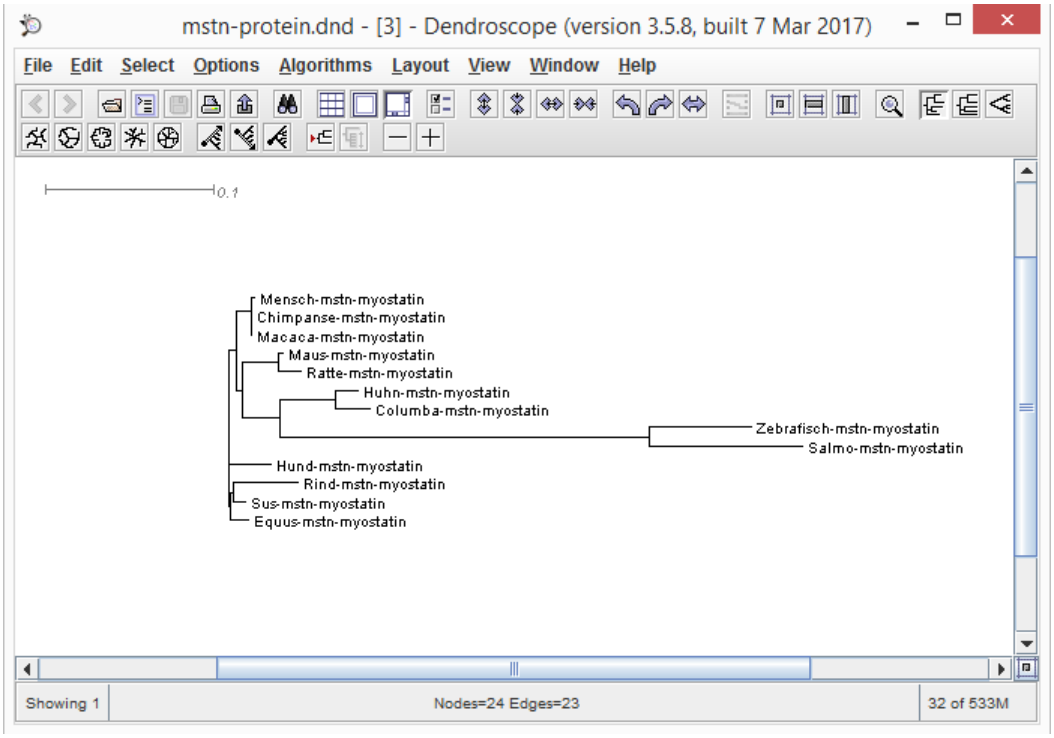
Beschäftigen Sie sich mit dem Ergebnis und prüfen Sie, ob der Baum mit Ihrer Vorstellung über den Verwandtschaftsgrad der Arten übereinstimmt. Damit die Aufgabe nicht zu leicht ist, sind zu den Arten nicht die deutschen Namen sondern die lateinischen Artnamen (in gekürzter Form) angegeben. Finden Sie heraus, um welche Arten oder Familien es sich bei *Equus*, *Sus*, *Macaca*, *Columba* und *Salmo* handeln könnte.

Das Ergebnis zeigt, dass man durch Vergleich von Proteinsequenzen Aussagen zum Stammbaum von Arten machen kann. Stammbäume werden heute üblicherweise mit Proteinsequenzen erstellt. Allerdings betrachten Wissenschaftler hierzu verschiedene Proteinfamilien und verwenden komplexere Computerprogramme, bevor die Stammbäume für die Einordnung von Arten in eine Gruppe als wissenschaftlich gesichert gelten können.

Das Ergebnis von Tour 1.4.2 ist hier für alle hinzugefügten Arten dargestellt:



Das Phylogramm hierzu sieht so aus:



Alternativ kann man diesen Stammbaum mit der oben markierten Ansicht darstellen lassen. Welche der beiden Ansichten übersichtlicher ist, ist reine Geschmackssache.

1.5 Vergleich der Stammbäume von zwei verwandten Proteinfamilien

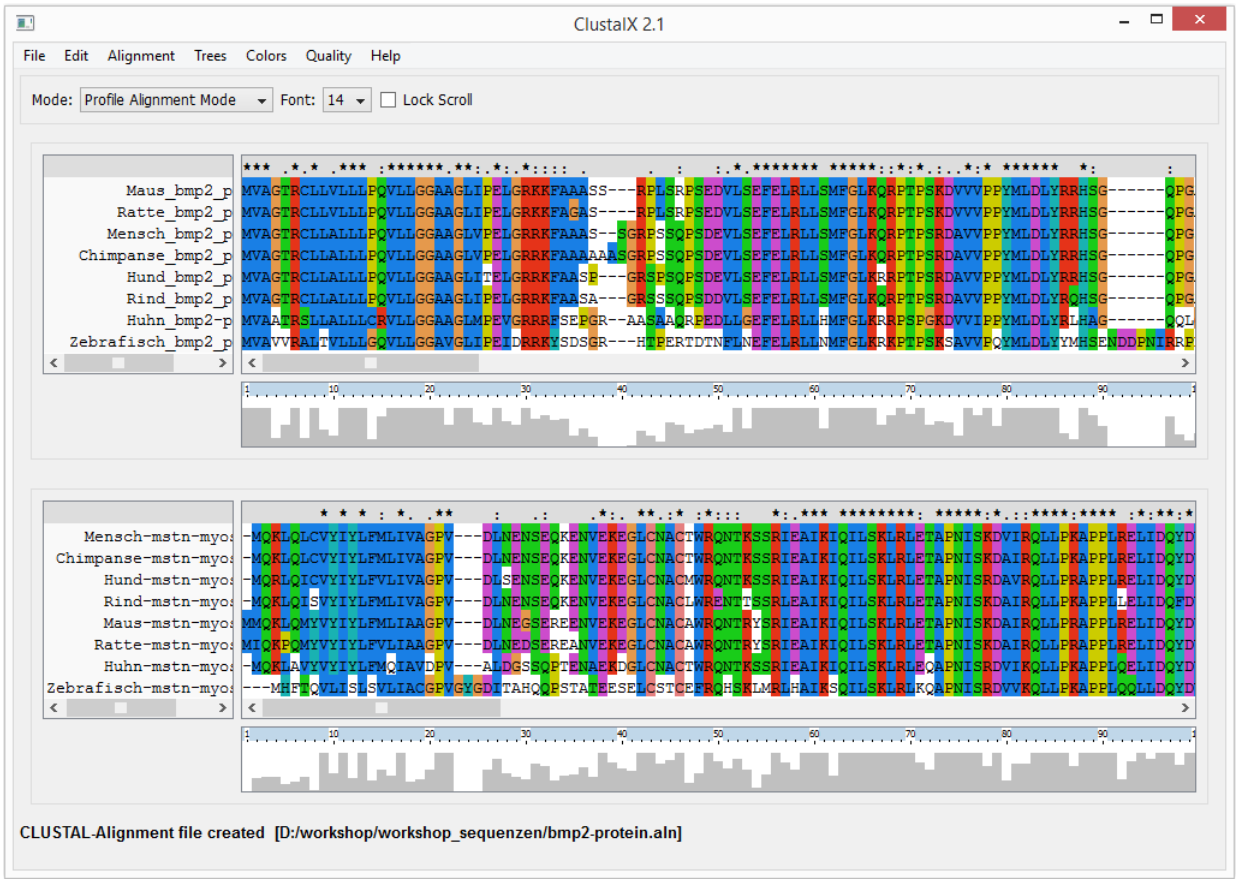
30-60 Min.

1.5.1 Für die Proteinfamilie *bmp2* ein MSA erstellen

Bmp2 ist eine mit dem Protein Myostatin eng verwandte Proteinfamilie. Wir erstellen ein MSA aus den Sequenzen in `bmp2-protein.fas`.

Zu Vergleich von MSAs auf *Profile alignment mode* klicken und dann das Alignment der Proteinfamilie *mstn* laden (Datei `mstn-protein.aln`). Wichtig: Auf die Dateiendung achten, `.aln` und nicht `.fas` wird benötigt.

Das Ergebnis sieht so aus:



Vergleich der beiden Proteinfamilien. Sind die Alignments so verschieden, dass man von zwei Familien sprechen kann?

Tipp: Am Ende des Alignments sind die Sequenzen besser zu vergleichen.

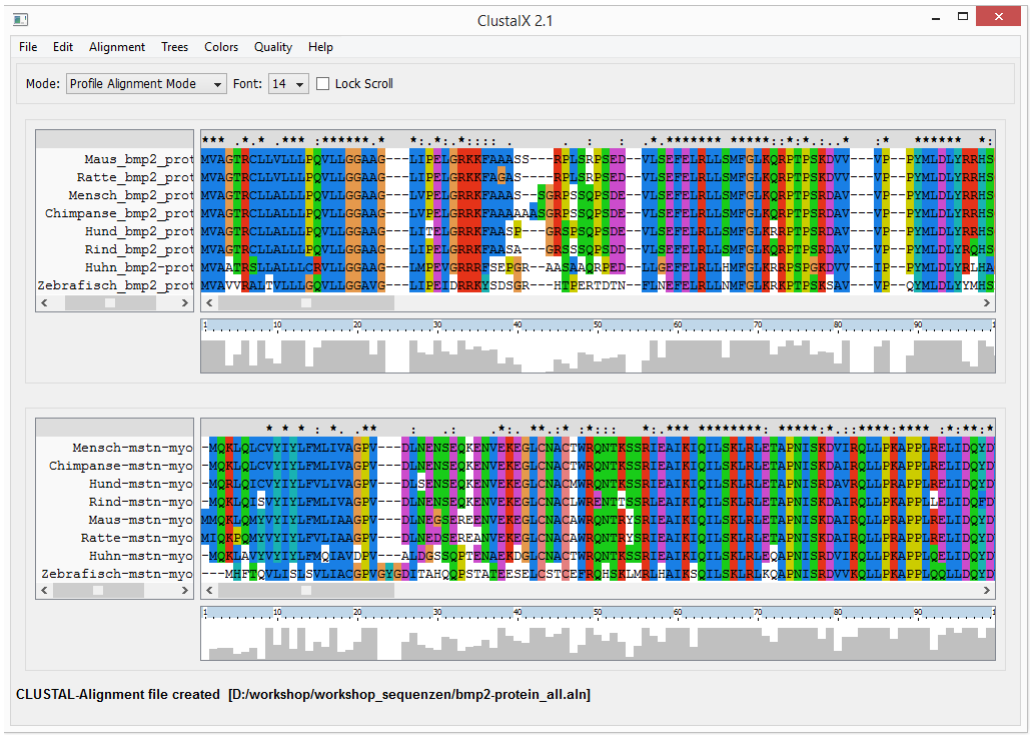
Die Frage kann man nicht 'mal eben so' beantworten. Dazu muss man schon Fachmann sein. Im Rahmen dieses Workshops kann man nur diskutieren und natürlich weitere Alignments erstellen, siehe Aufgabe 1.5.2.

1.5.2 Hinzufügen des MSA von *mstn* zum MSA von *bmp2*

Erstellen eines gemeinsamen Alignments durch Klicken auf *Alignment* und *Align profile 2 to profile 1*.

Anhand des Ergebnisses kann man besser beurteilen, in welchen Bereichen die Sequenzen sehr ähnlich sind und in welchen Bereiche sie sich unterscheiden.

Kann man die Lücken, die z.B. in *bmp2* zwischen Position 20 und 30 auftreten, erklären?

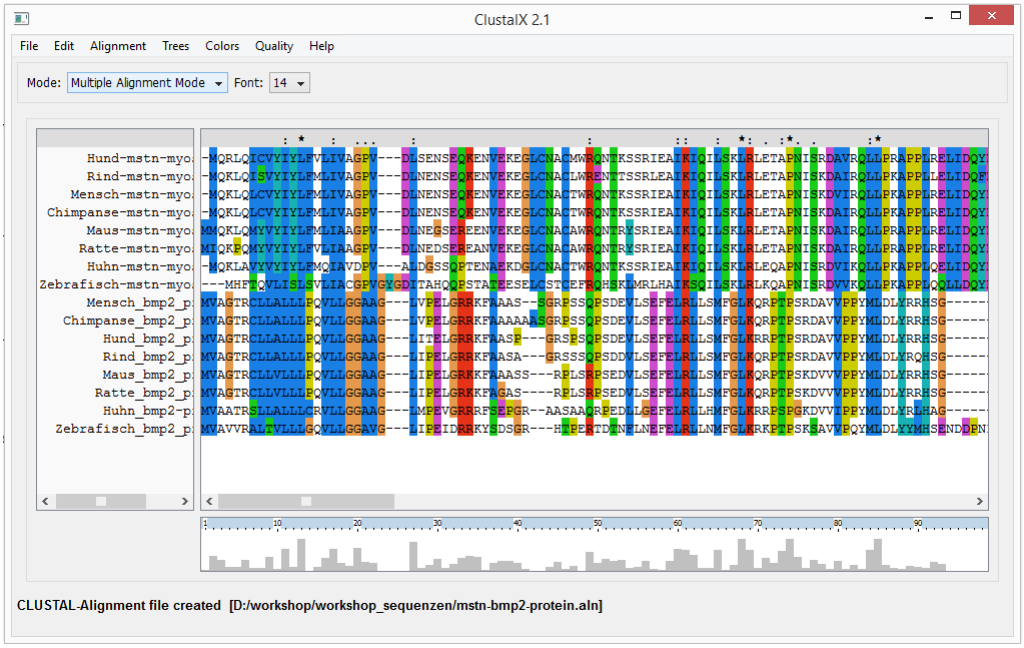


1.5.3 Erstellen eines gemeinsamen MSA der Sequenzen aus *mstn* und *bmp2*

Der Fachmann/die Fachfrau wäre mit dem Alignment aus Aufgabe 1.5.2 soweit zufrieden, würde sich aber die Frage stellen: Wie sieht ein Alignment aus, das in einem Schritt, also mit den Sequenzen von *mstn* und *bmp2* in einer Datei `mstn-bmp2-protein.fas` erstellt wird?

Tipp: Das Programm ClustalX beenden und neu öffnen, dann die Datei mit dem Namen `mstn-bmp2-protein.fas` laden und das Alignment erstellen.

Ergeben sich hier Änderungen oder werden die Sequenzen ähnlich gruppiert wie im *profile alignment*?



In Schritt 1.5.2 und Schritt 1.5.3 wurde versucht die zu klären, ob beide Proteinfamilien *mstn* und *bmp2* eng verwandt sind. Die Antwort hier im: ja.

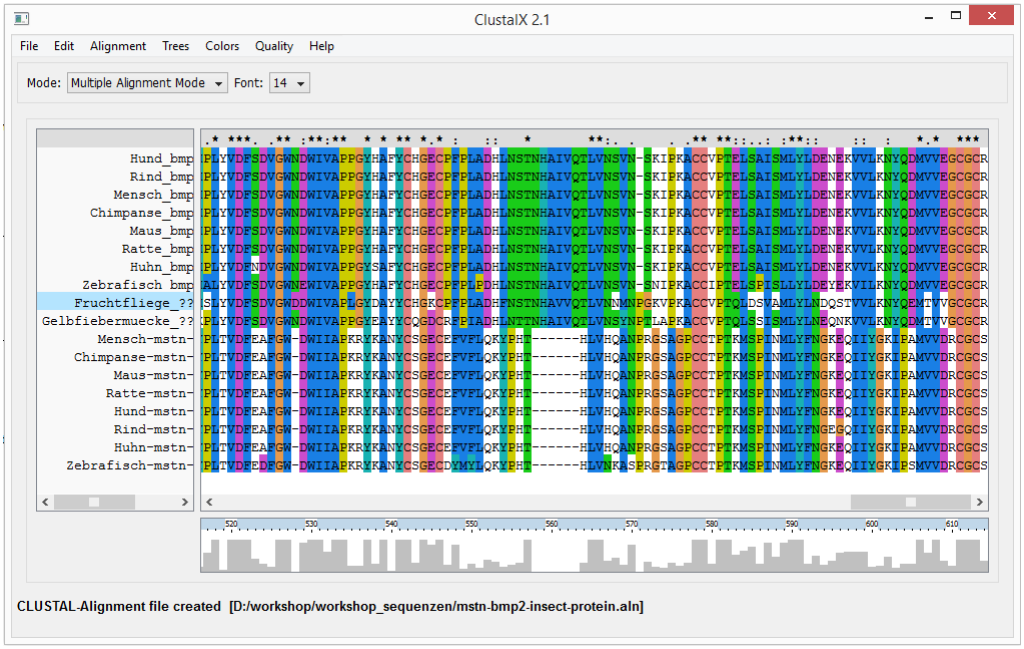
Es gehören sogar noch zwei weitere Sequenzen zu der Proteinfamilie von *mstn* und *bmp2*. Diese Sequenzen sollen im nächsten Schritt hinzugefügt werden.

1.5.4 Zum MSA der Proteinfamilien *mstn* und *bmp2* die Sequenzen der Fruchtfliege und der Gelbfiebermücke hinzufügen

Zunächst werden alle *gaps* (Lücken) entfernt, indem die Namen der Sequenzen markiert und *Edit/Remove all gaps* ausgewählt wird. Unter *File/Append sequence* werden dann die Dateien `Fruchtfliegeprotein.fas` und `Gelbfiebermuecke-protein.fas` hinzugefügt. Das Alignment wird über *Do complete alignment* erstellt.

Tipp: Für die Ergebnisdateien einen *neuen* Namen vergeben, also z.B. `mstn-bmp2-insect-protein.fas`. Ansonsten würden die Ergebnisdateien aus Aufgabe 1.5.3 überschrieben.

Gehören die Proteine der beiden Insektenarten zu einer der Proteinfamilien *mstn* oder *bmp2*, und wenn ja, zu welcher?



Eine alternative Ansicht bekommt man, wenn man den Leitbaum mit dem Programm Dendroscope öffnet.

2 Reiseführer

2.1 Darwin oder Evolution als Basis, um Biologie zu verstehen

Charles Darwin, der übrigens verschiedene andere Fächer studierte, bis er sich der Biologie zuwandte, hatte 1831–1836 die Gelegenheit, an einer Forschungsreise teilzunehmen. Mit dem Schiff „Beagle“ reiste er 5 Jahre rund um die Welt, sammelte und beobachtete Pflanzen und Tiere. Die genaue Auswertung dauerte ca. 10 Jahre und führte zu seiner Theorie der „Evolution durch natürliche Selektion“:

- Arten ändern sich im Laufe der Zeit.
- den Prozess der Änderungen beschreibt Darwin als „natürliche Selektion“ (im Gegensatz zur künstlichen Selektion, die in der Tier- und Pflanzenzucht bereits seit Jahrtausenden betrieben wird).

Hierbei konnte Darwin seine Theorie „nur“ auf den Vergleich von Arten (lebende Arten und Fossilien) begründen.

Individuen einer Art zeigen oft Merkmale (z.B. Größe), die u.a. in der Züchtung genutzt werden, um durch Vermehrung dieser Individuen diese Merkmale auszuprägen.

2.2 Wie funktioniert natürliche Selektion (nach Darwins Thesen)?

Organismen (Individuen) mit bestimmten Eigenschaften überleben und vermehren sich besser als andere Individuen in einer Population. Diese natürliche Selektion führt zu einer Adaption (Anpassung): Struktur-, physiologische und Verhaltenseigenschaften fördern das Überleben und die Reproduktion unter den jeweiligen Umweltbedingungen. Die unterschiedlichen Umweltbedingungen führen zu unterschiedlichen Anpassungen im Laufe der Erdgeschichte und zur Entwicklung der beeindruckenden Artenvielfalt.

Evolution, die Entstehung und Entwicklung des Lebens, können wir auch wie folgt zusammenfassen:

- Arten entwickeln sich aus Vorläuferarten durch Änderungen (Modifikationen).
- Alle Arten lassen sich also auf gemeinsame Vorfahren zurückführen (Stammbaum des Lebens).

Man kann dies auch auf Zellebene sehen:

- Alle Zellen entwickeln sich aus Zellen und sind aus einer Vorläuferzelle entstanden.
- Alle Zellen lassen sich auf eine Urzelle zurückführen.

Gleiches gilt auch für Biomoleküle, z.B. entstehen DNA und Proteine durch Evolution aus Vorläufermolekülen.

2.3 Was ist eine Art?

Unter Art versteht man (vereinfacht) eine Gruppe von Organismen, die ähnlich aussehen und miteinander Nachkommen erzeugen können. Nachkommen unterscheiden sich von ihren Eltern. Betrachtet man alle Individuen einer Art (also die Population), so finden sich viele Variationen.

2.4 Was Darwin nicht wissen konnte

Darwin (und Zeitgenossen, wie z.B. Gregor Mendel) konnten Evolution nur anhand von Phänotypen und deren Variation zwischen verschiedenen Arten beobachten. Seit 50 Jahren kennen wir den Genotyp (DNA) und können durch den Vergleich von Sequenzen Mutationen finden und diese oft einem Phänotyp zuordnen.

2.5 Was ist DNA und wie arbeitet man mit DNA am Computer?

DNA oder Desoxyribonukleinsäure (desoxyribonucleic acid) ist in der ersten Hälfte des 20. Jahrhundert als die Informationsquelle erkannt worden, die von Zelle zu Zelle, von Generation zu Generation, von Vorläuferart zu neu entstandener Art weitergegeben wird. Mit der Strukturaufklärung durch die Wissenschaftler Watson und Crick wurden bereits 1953 drei der vier wichtigen Eigenschaften der DNA vorgeschlagen.

DNA ist das Material, in dem die genetische Information eines Organismus gespeichert wird.

- die Information ist als Folge von sog. Basen oder Nukleotiden `A`, `C`, `G`, `T`, ohne Punkt und Komma, gespeichert sind, z.B.:

  ```
  GGATTCGAATCCTCTC...
  ```

  Man spricht daher auch von einer DNA-Sequenz.[3]
- Je nach Organismus sind diese DNA-Sequenzen unterschiedlich lang, z.B. bei Viren in der Grössenordnung von $10^4 - 10^5$, bei Bakterien in der Grössenordnung von $10^6 - 10^7$ und bei Wirbeltieren in der Grössenordnung von $10^9 - 10^{10}$.
- Die DNA-Sequenz unterscheidet sich zwischen Individuen (s.u.)

DNA kann durch Mutationen verändert werden. Z.B. kann eine Base `A` durch eine andere Base `C`, ersetzt werden. Diese Änderung ist permanent, d.h. sie wird an die Nachkommen weiter gegeben. Hier ein Beispiel:

Individuum 1	`....GGAACGTA...`
Individuum 2	`....GAAACGTA...`
Individuum 3	`....GGTACGTA...`

An Position 2 steht in der DNA von Individuum 1 ein `A` statt ein `G`, und in der DNA von Individuum 3 ist an Position 3 ein `T` statt ein `A` zu finden.

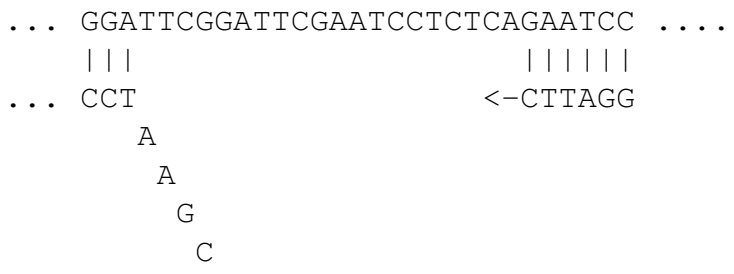
Die DNA wird vor der Zellteilung verdoppelt. Dieser Prozess heißt Replikation und läuft sehr präzise (fehlerfrei) ab. Für unseren Workshop betrachten wir die Replikation einmal sehr abstrakt und achten nur darauf, was mit der Information der DNA (der DNA-Sequenz) passiert. DNA ist ein Molekül, das aus einem Doppelstrang besteht:

[3] Tatsächlich sind die Basen in einer Doppelhelix gespeichert. Dieses erlaubt einerseits durch komplementärer Basen und Brückenbildung eine hohe Stabilität zu erreichen und andererseits eine räumliche Flexibilität, damit sich die DNA auf engem Raum verwinden kann. Dieser Aspekt der DNA soll aber hier keine Rolle spielen.

```
....GGATTCGAATCCTCTCA....
    ||||||||||||||||
....CCTAAGCTTAGGAGAGT....
```

`G` und `C` „paaren“, ebenso wie `A` und `T`. Man spricht von komplementären Basen.

Bei der Replikation entsteht nun aus dem Doppelstrang zwei Einzelstränge (Matrize, die „Kopiervorlage“) und zu jedem Strang wird ein neuer Strang synthetisiert (hier nur für den oberen Strang gezeigt):



So entstehen exakte Kopien der DNA. Wichtig für die Sequenzanalyse ist folgende Tatsache: Die DNA-Stränge haben eine Richtung.

Man sagt dann auch:

- Die DNA-Stränge sind antiparallel.
- Die DNA-Stränge sind revers komplementär.

Die DNA-Sequenz wird (fast) immer in der 5'→3'-Richtung notiert.

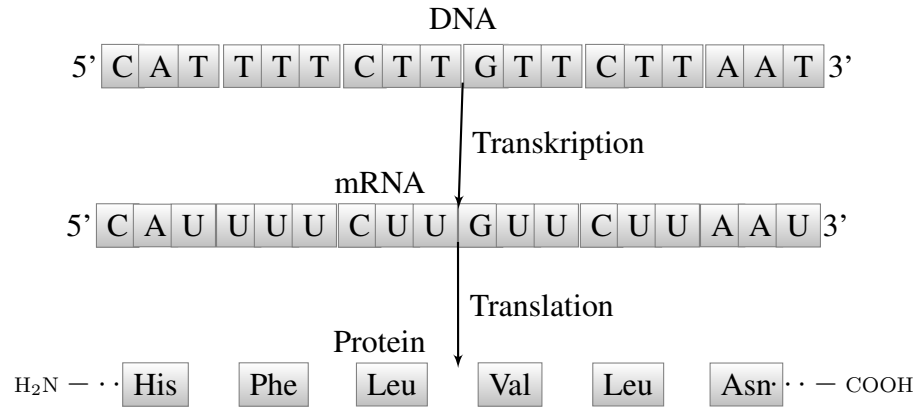
Das genetische Material (der Genotyp) exprimiert den Phänotyp oder (übersetzt in die bisher benutzten Fachworte): Die DNA-Sequenz (der Informationsträger) wird benutzt, um zum Beispiel Proteine aufzubauen, die maßgeblich den Phänotyp einer Zelle (eines Organismus) bestimmen.

zur Landkarte für Pauschalreisende, Abschnitt 1.

2.6 Was ist ein Gen?

In der DNA lassen sich Informationseinheiten nachweisen, die als Gene bezeichnet werden. Ein Gen ist eine Einheit, die die Information für ein Protein enthält. In einem als Transkription bezeichneten Prozess können diese Einheiten der DNA in Boten-RNA (messenger RNA oder mRNA) umgeschrieben werden. Diese Moleküle bringen genetische Informationen zu den Ribosomen. Dort wird in einem Prozess, den man Translation nennt, ihre Nukleotidsequenz in eine Aminosäuresequenz übersetzt. Das dabei entstehende Biomolekül ist ein Protein.

Kurz gefasst: DNA → mRNA → Protein, wie in der folgenden Abbildung beispielhaft dargestellt:



Dabei codieren je 3 Nukleotide (ein sog. Codon) für eine Aminosäure. Die Übersetzung (Translation) erfolgt entsprechend der folgenden Übersetzungstabelle, die sich je nach Organismus leicht unterscheiden kann.

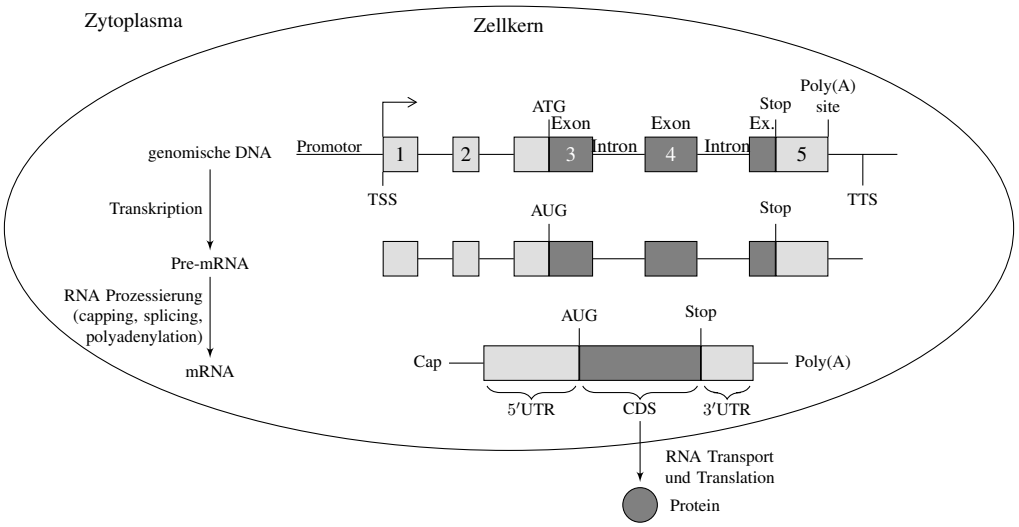
TTT	Phe	F	TCT	Ser	S	TAT	Tyr	Y	TGT	Cys	C
TTC	Phe	F	TCC	Ser	S	TAC	Tyr	Y	TGC	Cys	C
TTA	Leu	L	TCA	Ser	S	TAA	stop	*	TGA	stop	*
TTG	Leu	L	TCG	Ser	S	TAG	stop	*	TGG	Trp	W
CTT	Leu	L	CCT	Pro	P	CAT	His	H	CGT	Arg	R
CTC	Leu	L	CCC	Pro	P	CAC	His	H	CGC	Arg	R
CTA	Leu	L	CCA	Pro	P	CAA	Gln	Q	CGA	Arg	R
CTG	Leu	L	CCG	Pro	P	CAG	Gln	Q	CGG	Arg	R
ATT	Ile	I	ACT	Thr	T	AAT	Asn	N	AGT	Ser	S
ATC	Ile	I	ACC	Thr	T	AAC	Asn	N	AGC	Ser	S
ATA	Ile	I	ACA	Thr	T	AAA	Lys	K	AGA	Arg	R
ATG	Met	M	ACG	Thr	T	AAG	Lys	K	AGG	Arg	R
GTT	Val	V	GCT	Ala	A	GAT	Asp	D	GGT	Gly	G
GTC	Val	V	GCC	Ala	A	GAC	Asp	D	GGC	Gly	G
GTA	Val	V	GCA	Ala	A	GAA	Glu	E	GGA	Gly	G
GTG	Val	V	GCG	Ala	A	GAG	Glu	E	GGG	Gly	G

Aminosäuren werden meistens im 1-Buchstaben-Code geschrieben, weil dies nicht nur Platz spart, sondern auch übersichtlicher ist.

Da jede Aminosäure eine Seitenkette mit bestimmten physikalischen und chemischen Eigenschaften hat, ist die Abfolge der Aminosäuren in der Proteinsequenz maßgeblich für die Proteinfaltung und damit für die Proteinstruktur. Die Gestalt und Funktion von Zellen, und damit auch ganzer Organismen, wird im Wesentlichen von Proteinen bestimmt. Sie prägen in ihrer Gesamtheit den Phänotyp, also die charakteristischen morphologischen (äußerlich sichtbaren) Eigenschaften, ganz erheblich.

In allen Eukaryoten (Arten, die über einen Zellkern verfügen) sind Gene etwas komplexer aufgebaut. Die Gene bestehen im Genom (auf den Chromosomen) aus Exons und Introns. Bei

der Transkription wird zunächst eine sog. hnRNA erstellt, in der Exons und Introns in eine RNA kopiert sind. Anschließend werden die Introns aus der hnRNA entfernt. Der Prozess wird Spleißen genannt und führt über zwei weitere Schritte (5' Capping und 3' Polyadenylierung, die im Workshop nicht betrachtet werden) zu einem mRNA Molekül, in dem die Exons aneinandergereiht sind.



In Tour 1.3 wird das Thema Genaufbau etwas genauer beleuchtet. Abschnitt 2.13 befasst sich mit den Informationsfluss im Kontext von Genen.

2.7 Was ist ein Genom?

Die Gesamtheit der vererbbaren Informationen eines Organismus wird Genom genannt. In Wirbeltieren (Säugetiere, Reptilien, Fische und Vögel) ist das Genom auf mehrere Chromosomen verteilt.

Seit mehr als 20 Jahren kann man die DNA-Sequenzen, also die Nukleotidabfolgen, von Genomen ermitteln und hat somit die gesamte Information zur Verfügung. So wurde beispielsweise im Jahr 1995 das erste bakterielle Genom sequenziert und im Jahre 2001 die vollständige Sequenz des menschlichen Genoms entschlüsselt. Dieser Fortschritte verleitete den damaligen US-Präsidenten Clinton zu der Bemerkung:

> „Wir können in unserem Genom lesen wie in einem Buch".

Nun, Recht hatte er! Allerdings ist das Buch in einer Fremdsprache geschrieben, aber wir verstehen immer besser die Grammatik und viele neue Vokabeln. Wer ist „wir"? Nun, sicherlich mehrere tausend Wissenschaftler, die täglich im Labor und am Computer neue Eigenschaften und Informationen in Genomen entdecken.

2.8 Wo findet man DNA- und Proteinsequenzen?

Die meisten entschlüsselten Genome werden in Datenbanken gespeichert, die über das Internet frei für jedermann zugänglich sind. So können Wissenschaftler genomische Daten nutzen, um ihre speziellen Fragestellungen zu bearbeiten. Wir nutzen in diesem Workshop diese Daten und sehen uns dabei in der Welt der Gene und Genome um.

Tour 1.1→1.2→1.3 zeigt wie den Weg eines Gen von der DNA über die mRNA bis zum Protein.

Tour 1.2→1.4→1.5 zeigt wie man Stammbäume mit Proteinfamilien erstellt.

Die notwendigen Sequenzdaten haben wir bereits herausgesucht und in verschiedenen Dateien zusammengestellt. Diese Dateien findet man unter `http://www.zbh.uni-hamburg.de/` im Bereich Studium→Informationen für Schüler/-innen→ Workshop DNA-Sequenzanalyse.

2.9 Wie kann man eigene Datensätze aus öffentlichen Datenbanken recherchieren?

Natürlich kann man sich auch direkt in Datenbanken mit DNA- und Proteinsequenzen umsehen oder Programme nutzen, die diese Daten analysieren, sowohl im Web-Browser als auch offline im Desktop.

- Es handelt sich um Informationsquellen, die „von Wissenschaftlern für Wissenschaftler“ erstellt wurden. Das bedeutet, dass vielfach das Hintergrundwissen fehlt, um mit diesen Informationen etwas anzufangen.
- Also halten wir uns an die Landkarte für Pauschalreisende.
- Die Informationsquellen rund um Gene, Genome und Proteine sind überwiegend in englischer Sprache. Englisch ist nun einmal die Sprache der Wissenschaft. Wer Englisch in der Mittelstufe hatte, sollte das meiste problemlos lesen können. Nur die Fachworte sind natürlich schwieriger.

Tipp: Intuitiv lesen, man muss nicht jedes Wort verstehen.

2.10 Stammbaumanalyse

Auf Basis der Unterschiede zwischen den DNA- oder Proteinsequenzen aus Genen gleichen Ursprungs (sogenannte homologe Gene, die sich auf einen gemeinsamen Vorfahren zurückführen lassen), kann man Untersuchungen über die Verwandtschaftsverhältnisse zwischen Arten anstellen. Dazu vergleicht man die Sequenzen untereinander, da man davon ausgeht, dass sich bei weiter entfernt verwandten Arten im Laufe der Evolution mehr Sequenzunterschiede angesammelt haben. Haben zwei Sequenzen eine starke Ähnlichkeit, dann besteht eine hohe Wahrscheinlichkeit, dass beide Sequenzen homolog sind, da sie aus einem gemeinsamen Vorläufer-

Molekül entstanden sind. Beide Sequenzen sind also im evolutionären Sinne miteinander verwandt.

Die Phylogenie (griechisch „phylon“ = Stamm und „genia“ = Ursprung) bezeichnet die Schlussfolgerung von Sequenzunterschieden auf evolutionäre Verwandtschaftsverhältnisse. Der Begriff wurde um 1900 von Ernst Haeckel, einem Pionier der Evolutionstheorie, begründet. Das Ziel einer phylogenetischen Untersuchung ist es, entweder eine Gruppe von Organismen, eine Gruppe von DNA-Sequenzen oder von Proteinen auf ihre Verwandtschaftsverhältnisse zu untersuchen. Der Arbeitsablauf in der Forschung sieht dabei meist wie folgt aus:

1. Suche nach geeigneten Sequenzabschnitten (homologen Sequenzen, die sich aber genügend unterscheiden, um Aussagen treffen zu können),
2. Erstellen eines Datensatzes von solchen geeigneten Sequenzen aus den Genomen mehrerer Organismen,
3. Erstellung eines sogenannten multiplen Sequenzalignments (s.u.),
4. Entscheidung für ein geeignetes evolutionäres Modell zur Stammbaumerstellung,
5. Phylogenetische Analyse, d.h. Berechnung des Stammbaums,
6. Visualisierung des Stammbaums (als Grafik darstellen),
7. Statistische Analyse des Stammbaums, um seine Aussagekraft zu bestimmen,
8. Je nach Ergebnis von Schritt 7 wird ggf. eine Wiederholung von Schritt 2-6 vorgenommen. Insbesondere Schritt 5 wird ggf. mit anderen Methoden wiederholt, bis die geeignete Methode gefunden wurde.

Für die Stammbaumanalyse in diesem Workshop haben wir die Punkte 1 und 2 bereits vorab erledigt. Sie beginnen also mit Punkt 3. Für die Punkte 4 und 5 wird nur ein Modell und eine Methode benutzt, damit ausreichend Zeit bleibt, die Ergebnisse auszuwerten und zu diskutieren.

zur Landkarte, Tour 1.2.

zur Landkarte, Tour 1.4.

2.11 Wie kann man Unterschiede zwischen Sequenzen messen und vergleichen?

Die Antwort der Bioinformatik auf diese Frage sind Sequenzalignments. Bei einem Alignment versucht man, die Zeichen von zwei Sequenzen so auszurichten, dass die Reihenfolge der Zeichen erhalten bleibt und jedes Zeichen in der einen Sequenz einem Zeichnen in der anderen Sequenz oder oder einer „Lücke“ zugeordnet ist. So könnte ein paarweises Alignment zwischen den beiden DNA-Sequenzen `ATTGGGTATG` und `GTGTGGGTAAATGGT` folgendermaßen aussehen:

```
-ATTGGGT--ATG-
```

```
    |||||   |||
GTGTGGGTAAATGGT
```

Eine Fehlpaarung (zwei unterschiedliche gegenüberstehende Basen) in dem Alignment entspricht einer Mutation, in der ein Zeichen gegen einen anderen ausgetauscht ist. Die Lücken (gaps, dargestellt durch ein $-$) hingegen weisen auf eine Deletion (Löschung) oder eine Insertion (Einfügung) hin. Die einander zugeordneten (man sagt auch „alignierten“) Basen sollten identisch oder möglichst ähnlich sein, weil viele gleiche oder ähnliche Basen in gleicher Reihenfolge, wie oben beschrieben, auf eine evolutionäre Verwandtschaft hinweisen. Indem man für gleiche bzw. mutierte Zeichenpaare Punkte verteilt (z.B. $+1$ für zwei gleiche Basen an der gleichen Stelle, -3 für zwei verschiedene Basen, usw.), kann man nun eine Bewertung der Ähnlichkeit zweier Sequenzen angeben. Zur Hervorhebung von identischen Zeichenpaaren in den Sequenzen wird oben zwischen den alignierten Sequenzen das Symbol verwendet.

Die hier beschriebenen Prinzip von Alignments kann man auf mehr als zwei Sequenz erweitern. Man spricht dann von einem multiplen Sequenzalignment (MSA). Ein MSA ist eine Sammlung von drei oder mehr Protein- oder DNASequenzen, die teilweise oder über die gesamte Länge aligniert sind. Sequenzen werden zeilenweise gegenübergestellt und gleiche Positionen erscheinen in einer Spalte, wie im folgenden Beispiel mit DNA-Sequenzen:

Mensch	G	C	G	G	A	A	C	G	A	A	G	C	G	T
Huhn	G	C	G	A	C	G	G	A	–	C	A	G	G	T
Maus	G	C	G	A	A	T	C	–	–	G	A	G	G	T
Schwein	G	C	C	A	G	T	–	T	–	T	A	G	T	T

Während beim paarweisen Alignment die Suche nach verwandten (homologen) Sequenzen das Ziel ist, wird das MSA meist mit einem Datensatz durchgeführt, von dem bereits bekannt ist, dass es sich um verwandte Sequenzen handelt. Das MSA dient der Darstellung der Gemeinsamkeiten und Unterschiede. An dem Bild oben kann man z.B. sehr gut erkennen, dass sich etwa die Sequenzen der letzten beiden Organismen stärker von den anderen sechs, aber auch untereinander unterscheiden.

ClustalX ist das bekannteste Programm zur Berechnung von MSAs. Wir nutzen es auch auf unserer Entdeckungsreise. Es arbeitet schrittweise: Im ersten Schritt werden alle möglichen paarweisen Alignments zwischen jeweils zwei Eingabesequenzen erstellt, etwa so für 4 Sequenzen:

Sequenz 1 und Sequenz 2
Sequenz 1 und Sequenz 3
Sequenz 1 und Sequenz 4
Sequenz 2 und Sequenz 3
Sequenz 2 und Sequenz 4
Sequenz 3 und Sequenz 4

Basierend auf paarweisen Alignments wird dann ein sog. Leitbaum (engl. *guide tree*) erstellt, der beschreibt, in welcher Reihenfolge die paarweisen Alignments zusammengefügt werden.

ClustalX beginnt mit dem paarweisen Alignment des ähnlichsten Sequenzpaars, erweitert dieses dann jeweils mit den nächstverwandten Sequenzen. Das Verfahren endet, wenn alle Sequenzen in das MSA eingeflossen sind. ClustalX arbeitet sehr effizient und kann MSAs für viele, auch lange Sequenzen berechnen.

zur Landkarte, Tour 1.1.1.

zur Landkarte, Tour 1.1.2.

2.12 Was ist *Myostatin*?

Das Gen *Myostatin* wurde erstmals 1997 in der Maus entdeckt und in Zusammenhang mit einem Phänotyp beschrieben: Mäuse, deren *Myostatin*-Gen defekt ist, zeigen ein starkes Muskelwachstum, das auch als „Mighty Mouse“ bezeichnet wird.

Myostatin ist ein Protein (Eiweiß), das im menschlichen oder tierischen Körper gebildet wird. Es hat einen negativ regulierenden Einfluss auf das Muskelwachstum, das heißt es sorgt dafür, dass die Muskeln kontrolliert wachsen. Ohne das *Myostatin* würde Muskelgewebe einfach über das natürliche Maß weiterwachsen. Diese wichtige Funktion ist ein Hinweis darauf, dass *Myostatin* in vielen Lebewesen zu finden sein muss, was z.B. in der Tierzucht von Interesse ist.

Inzwischen wurde das Gen auch im Menschen gefunden und es wird untersucht, ob und ggf. wie *Myostatin* bei Erkrankungen, die zu Muskelschwäche führen eine Rolle spielt. Möglicherweise ließe sich ein Wirkstoff (Medikament) entwickeln der *Myostatin* hemmt (quasi den „Mighty Mouse“-Phänotyp erzeugt) und so die Erkrankung lindern, verzögern oder auch stoppen kann.

Es gibt aber auch Befürchtungen, dass solch ein Wirkstoff für Doping missbraucht werden kann. Doch Vorsicht! Derzeit ist zwar recht gut untersucht, welchen Einfluss *Myostatin* auf das Muskelwachstum hat, aber wenig zu seiner Funktion in der Haut, im Bindegewebe, im Gehirn und in weiteren Organen, in denen es auch gebildet wird. Ein Wirkstoff, der *Myostatin* hemmt, könnte also neben Muskelwachstum noch ganz andere Funktionen stören und so zu schweren Nebenwirkungen führen.

zur Landkarte für Pauschalreisende, 1.

zur Landkarte, Tour 1.3.

2.13 Aufbau eines Gens und Informationsfluss in der Zelle

Hier wird der Aufbau eines Gens etwas genauer beschrieben, damit die Ergebnisse aus Tour 1.3 verständlich sind.

Wie bereits Abschnitt 2.6 erläutert wurde, kann man den Prozess der Transkription und Translation auf Sequenzebene wie folgt beispielhaft beschreiben:

```
DNA      GATCTTTTATTATTTGAACTAAAAAAGAATTAATAAAACTTTCATTAATATGATCCA
                                  ↓ Transkription ↓

mRNA     GAUCUUUUAUUAUUUGAACUAAAAAAGAAUUAAUAAAACUUUCAUUAAUAUGAUCCA

                                  ↓ Translation ↓

Protein AspLeuLeuLeuPheGluLeuLysLysAsn******AsnPheHis***TyrAspPro
Protein D   L   L   L   F   E   L   K   K   N   *   *   N   F   H   *   Y   D   P
```

Für den Workshop sind folgende Dinge zu beachten:

1. Der Einfachheit halber sind mRNA-Sequenzen in der Regel im DNA-Code, also mit einem `T` (Thymin) statt einem `U` (Uracil) als Base geschrieben.
2. Proteinsequenzen können mit 3 Buchstaben abgekürzt oder auch im 1-Buchstaben-Code geschrieben werden. Meistens finden wir den 1-Buchstaben-Code.
3. Die mit einem $*$ markierten Stellen in der Proteinsequenz kennzeichnen Stop-Codons, die zu einem Abbruch der Translation führen.

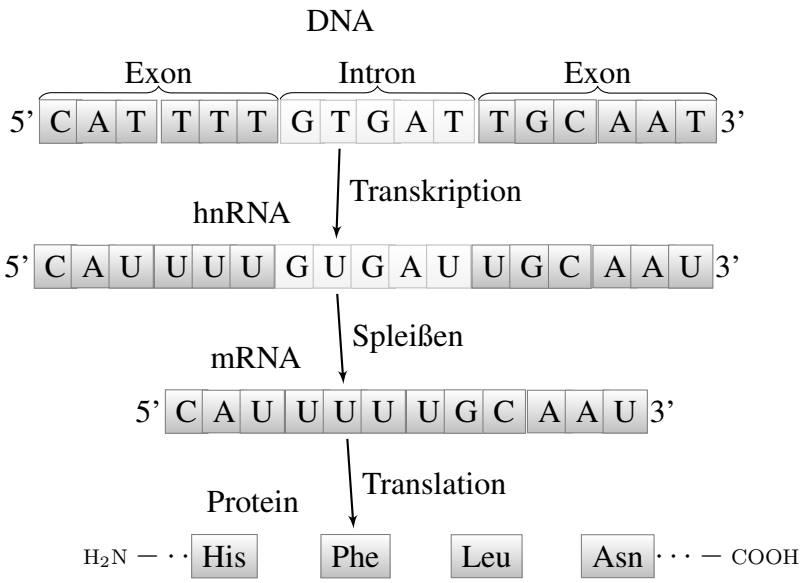
Gene werden in der Regel durch Sequenzieren der DNA oder mRNA ermittelt. Proteine bestimmt man selten direkt, weil das sehr aufwändig ist. Stattdessen leitet man die Aminosäuresequenz mit Hilfe von Computerprogrammen aus Gensequenzen ab. Dabei muss man zwei Dinge beachten:

1. Da immer 3 aufeinanderfolgende Basen (also ein Codon) für eine Aminosäure codieren, gibt es auch 3 verschiedene Leserahmen (engl. *reading frames*) auf dem Vorwärtsstrang.

```
      5′ GATCTTTTATTATTTGAACTAAAAAAGAATTAATAAAACTTTCATTAATATGATCCATCC 3′
         CTAGAAAATAATAAACTTGATTTTTTCTTAATTATTTTGAAAGTAATTATACTAGGTAGG
Rahmen 1 D  L  L  L  F  E  L  K  K  N  *  *  N  F  H  *  Y  D  P  S
Rahmen 2  I  F  Y  Y  L  N  *  K  R  I  N  K  T  F  I  N  M  I  H
Rahmen 3   S  F  I  I  *  T  K  K  E  L  I  K  L  S  L  I  *  S  I
```

Man muss also mit Hilfe von Programmen zunächst den richtigen Leserahmen ermitteln, um die Proteinsequenz zu bestimmen.

2. In allen Eukaryoten (Arten, die über einen Zellkern verfügen) sind Gene komplexer aufgebaut als oben beschrieben. Eukaryotische Gene bestehen im Genom (auf den Chromosomen) meist aus mehreren Exons und Introns. Der Informationsfluß von den Genen zu den Proteinen sieht so aus:



Bei der Transkription wird zunächst eine sog. hnRNA erstellt, in der Exons und Introns in eine RNA kopiert sind. Anschließend werden die Introns aus der hnRNA entfernt. Der Prozess wird Spleißen genannt und führt über zwei weitere Schritte (5’-Capping und 3’-Polyadenylierung), die im Workshop nicht betrachtet werden, zu einem mRNA Molekül, in dem die Exons aneinandergereiht sind.

In Tour 1.3 werden die (genomischen) DNA-Sequenzen mit den mRNA-Sequenzen verglichen, um die Exon-Intron-Struktur sichtbar zu machen. Es gibt zwei Bereiche, die nur in der genomischen DNA, nicht aber in der mRNA vorkommen. Das sind die Introns. Vergleicht man nun die Konservierung, d.h. wie viele Sequenzänderungen in Form von Nukleotidaustauschen (englisch mismatch) und die Anzahl von Lücken (englisch gaps), so sind Exons stärker konserviert als die benachbarten Intronbereiche. Wie kann man das erklären?

Man kann davon ausgehen, dass die Häufigkeit der Mutationen in Exons und Introns gleich hoch ist. Unterschiedlich ist jedoch die Selektion der Mutationen, was dann zu ihrer unter-

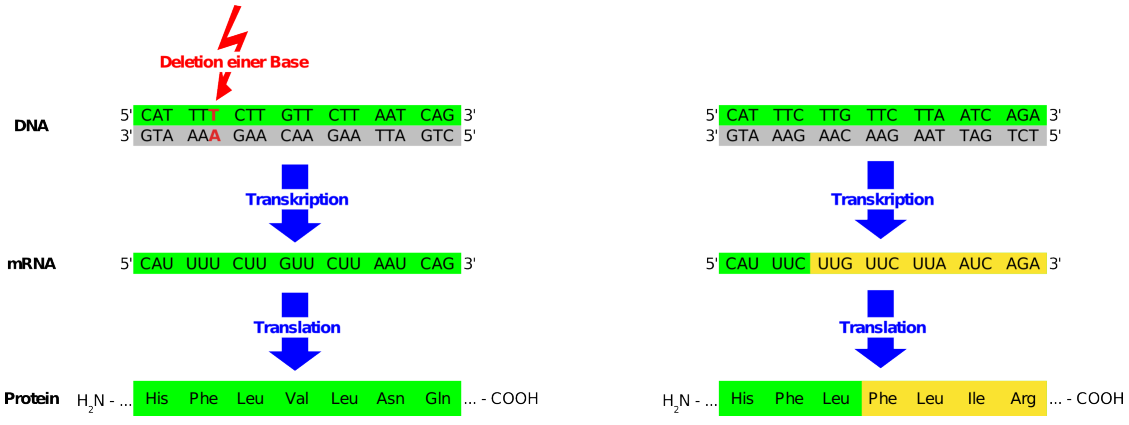
schiedlichen Konservierung führt. Jede Mutation kann folgende Effekte für das betroffene Individuum haben:

- Selektionsvorteil (sehr selten),
- Selektionsnachteil (häufig),
- neutral, d.h. keinen Selektionsvorteil oder Selektionsnachteil (häufig).

Mutationen mit Selektionsvorteil werden sich in einer Population durchsetzen, Mutationen mit Selektionsnachteilen gehen meist in wenigen Generationen verloren, weil die betroffenen Individuen sich nicht so erfolgreich reproduzieren können („survival of the fittest“). Neutrale Mutationen bleiben länger in einer Population erhalten. Dieses Zusammenspiel von Mutation und Selektion beobachten wir beim Vergleich von DNA- und mRNA-Sequenzen: Exons, aus denen sich die Proteinsequenzen ergeben, und die somit eine Funktion haben, sind stärker konserviert als Introns, die selten eine direkte Funktion ausüben.

Kurz gefasst, durch Selektion werden Mutationen aus DNA Abschnitten entfernt, die entscheidend zu einer Funktion beitragen. In der Regel sind Exons daher stärker konserviert als Introns.

Neben den genannten Unterschieden bzgl. der Konservierung gibt es weitere wichtige Eigenschaften von Exons und Introns: So sind insgesamt weniger Lücken in Exons als in Introns vorhanden und die Lücken haben oft eine Länge, die ein Vielfaches von 3 ist (eng benachbarte Lücken sollten hier zusammen betrachtet werden). Wie kommt solch ein Muster zustande? Nun, je 3 Nukleotide bilden ein Codon, d.h. codieren für eine bestimmte Aminosäure. Wenn 3 Nukleotide fehlen, fehlt im Protein nur eine Aminosäure, was möglicherweise nicht die Funktion des Proteins stört. Wenn aber 1 oder 2 Nukleotide fehlen, dann spricht man von einer sog. Leserahmen-Mutation, denn nach der Lücke wird in einem anderen Leserahmen weiter gelesen, was zu einer stark veränderten Proteinsequenz führt. Im Schema unten sind eine DNA-Sequenz und die resultierende Proteinsequenz abgebildet. Findet nun eine Deletion statt, dann sind die nachfolgenden (3’ gelegenen) Codons anders aufgebaut (gelber Bereich im Schema des Proteins rechts).



Dadurch ändert sich die anschließende Aminosäuresequenz und damit die Struktur des Proteins. Solche Proteine können oft nicht mehr ihre ursprüngliche Funktion in der Zelle ausüben.

Dies ist fast immer ein erheblicher Selektionsnachteil. Folglich finden wir Deletionen und Insertionen, die den Leserahmen verändern, nur sehr selten.

Übrigens ist dies auch der Grund, warum das Spleißen sehr präzise ablaufen muss. Introns müssen exakt an einer bestimmten Stelle herausgeschnitten werden, damit die zusammengefügten Exons dann auch einen sog. „Offenen Leserahmen“ enthalten.

zur Landkarte, Ende Tour 1.3.

2.14 Genstrukturvorhersage

Angenommen, die DNA-Sequenz eines bestimmten Gens auf einem Chromosom eines bestimmten Lebewesens ist bereits bekannt. Als nächstes interessiert sich der Forscher für den internen Aufbau dieses Gens. Das heißt, er möchte wissen, wie viele Exons und Introns es gibt, wo sie im Gen zu finden sind (also wo sie beginnen und enden) und wie lang sie sind. Diese Information ist wichtig, um zu wissen, welche Bereiche der Sequenz aus der mRNA vor dem Umschreiben in Protein herausgespleißt werden. Ohne diese Information würde man, wenn man die Sequenz mit dem genetischen Code übersetzt, falsche Aminosäuresequenzen bekommen.

Eine durchgehende DNA-Sequenz nennt man einen offenen Leserahmen (englisch open reading frame (ORF)), wenn sie aus einer Folge von Codons besteht, dessen erstes Codon ein Startcodon ist (meist `ATG`) und dessen letztes Codon ein Stopcodon (meist `TAG`, `TAA` oder `TGA`) ist. Der Name kommt daher, dass ein solcher ORF während der Translation komplett durchgelesen werden kann und als Vorlage für ein fertiges Protein dient. Die Vorhersage von Genen (mit den genauen Exon- und Intronpositionen) ist z.B. eine typische Aufgabe, mit der sich Bioinformatiker beschäftigen.

zur Landkarte, Ende Tour 1.3.

2.15 Sequenzen speichern und Sequenzformate

Daten am Computer, seien es Texte, Photos oder Filme werden in verschiedenen Dateiformaten gespeichert. Auch für biologische Daten gibt es spezielle Dateiformate. Im Workshop werden wir 3 solche Formate kennen lernen: Das FASTA Format, das Alignment-Format und ein Format für Leitbäume.

zur Landkarte, Start Tour 1.1.

2.15.1 Das FASTA Format

Das FASTA-Format ist ein einfaches, weit verbreitetes Dateiformat für biologische Sequenzen. Die Endung ist nicht fest vorgeschrieben, aber meist wird `.fasta`, `.fas`, oder `.fa` verwendet. Das FASTA-Format ist ein Textformat und kann mit vielen verschiedenen Programmen, z.B. Texteditoren, geöffnet werden. Die erste Zeile einer Sequenz in einer FASTA

Datei beginnt mit dem Zeichen >. Die Zeile enthält eine kurze Beschreibung der Sequenz (etwa: eindeutiger Identifikator, Organismus, Name des Gens oder Proteins, usw.). In der nächsten Zeile folgt die Basen- oder Aminosäuresequenz, die an beliebiger Stelle umgebrochen werden kann. Zum Beispiel sieht der Anfang einer FASTA Datei für *Myostatin*-Gen der Maus so aus:

```
>Maus_Mstn_gene_region
ATGCAAAAACTGCAAATGTATGTTTATATTTACCTGTTCATGCTGATTGC
TGCTGGCCCAGTGGATCTAAATGAGGGCAGTGAGAGAGAAGAAAATGTGG
AAAAAGAGGGGCTGTGTAATGCATGTGCGTGGAGACAAAACACGAGGTAC
TCCAGAATAGAAGCCATAAAAATTCAAATCCTCAGTAAGCTGCGCCTGGA
AACAGCTCCTAACATCAGCAAAGATGCTATAAGACAACTTCTGCCAAGAG
```

Eine FASTA-Datei kann auch mehrere solcher Sequenzeinträge enthalten. Diese werden dann einfach hintereinander geschrieben. Das FASTA-Format wird vom Programm ClustalX als Eingabeformat akzeptiert, d.h. die Sequenzen werden vom Programm angezeigt und weiterverarbeitet.

zur Landkarte, Tour 1.1.1.

2.15.2 Das Clustal-Format

Das Programm ClustalX erstellt aus Sequenzen ein MSA und speichert es in einer neuen Datei ab. Die neue Datei hat die Endung `.aln`. Öffnet man die Datei in ClustalX, so erhält man das Alignment angezeigt. Öffnet man die Datei mit einem Editor (z.B. WordPad unter MS Windows, TextEdit unter Mac OS X oder KWrite unter Linux), dann kann man die Sequenzen erkennen, allerdings mit den durch das Programm eingefügten Lücken.

2.15.3 Das Newick-Format für den Leitbaum

ClustalX gibt neben der Alignment-Datei auch eine weitere Ergebnisdatei mit der Endung `.dnd` aus. Diese Datei enthält die Informationen für den errechneten Leitbaum. Eng verwandte Sequenzen sind in Klammern gefasst und die Astlängen des Baums sind mit den jeweils errechneten Abständen versehen. Diese Datei wird z.B. vom Programm Dendroscope eingelesen und graphisch dargestellt. Alternativ kann man sich die Dateien mit der Endung `.dnd` auch unter `http://iubio.bio.indiana.edu/treeapp/treeprint-form.html` hochladen und darstellen lassen.

zur Landkarte, Tour 1.1.2.

2.15.4 Das Genbank-Format

Das Genbank-Format, ein weiteres gängiges Format für biologische Sequenzdaten, enthält weitere Daten, wie Informationen darüber, wann, von wem und wie die Sequenz das erste

Mal veröffentlicht wurde, welche besonderen Regionen (etwa Exons/Introns usw.) an welcher Stelle der Sequenz enthalten sind.

zu Landkarte für Pauschalreisende, 1.

2.16 Kurze Einführung in das Programm ClustalX

Die Installation des Programms ist nicht notwendig, wenn Sie den Workshop an den Rechnern im Zentrum für Bioinformatik durchführen. ClustalX ist kostenlos für die Betriebssysteme MS Windows, Mac OS X und Linux unter der URL `http://www.clustal.org/download/current` verfügbar. Für den Workshop verwenden wir ClustalX version 2.1.

Achtung: Nicht ClustalW, sondern ClustalX speichern. Durch einen Doppelklick auf die für Ihr Betriebssystem passende Version öffnet sich das Installationsprogramm.

Zunächst sollte man das Programm starten und sich einen Überblick verschaffen. Tour 1.1.1 in der Landkarte für Pauschalreisende gibt hier einen Überblick.

Generell gilt:

- Das Programm erwartet als Dateien mit Sequenzen im FASTA-Format (Dateiendung `.fas`, `.fasta` oder `.fa`).
- Das Programm zeigt das Alignment direkt an, die dazugehörige Datei hat die Endung `.aln`. Zusätzlich wird der Leitbaum ausgegeben (Dateiendung `.dnd`), der mit dem Programm Dendroscope angesehen werden kann.
- Die Dateien mit der Endung `.aln` können mit ClustalX angesehen werden.
- Sollten die vier verschiedenen Basen nicht in 4 verschiedenen Farben angezeigt sein, dann das Menü *Color/Load Color Parameter File* auswählen. Hier die Datei `coldna.xml` (bzw. für Proteine `colprot.xml`) laden. Diese befinden sich in dem Ordner, in dem ClustalX installiert ist.

zur Landkarte für Pauschalreisende, Abschnitt 1.

2.17 Kurze Einführung in das Programm Dendroscope

Die Installation des Programms ist nicht notwendig, wenn Sie den Workshop an den Rechnern im Zentrum für Bioinformatik durchführen.

Eine Übersicht zum Programm Dendroscope findet man unter

`http://www-ab.informatik.unituebingen.de/software/dendroscope`,

Hier gibt es auch ein Benutzerhandbuch. Dendroscope gibt es kostenfrei für die Betriebssysteme MS Windows, Mac OS X und Linux unter

`http://www-ab.informatik.uni-tuebingen.de/data/software/dendroscope/download/welcome.html`

Durch einen Doppelklick auf die für Ihr Betriebssysteme passende Version öffnet sich das Installationsprogramm. Hier die Installationsschritte befolgen.

Bei der Verwendung von Dendroscope beachten Sie bitte folgendes:

- Das Programm erwartet die Eingabedatei im sog. Newick-Format (Dateiendung `.dnd`).
- Wenn nicht alle sondern nur einige Arten zum Stammbaum hinzugefügt werden sollen, dann können ggf. Arten durch Klicken auf einzelne Sequenzen unter dem Menü *Edit/Clear sequence selection* entfernt werden, bevor das MSA und der Leitbaum erstellt werden.
- Alternativ kann man sich die Dateien mit der Endung `.dnd` auch über die Internetseite `http://iubio.bio.indiana.edu/treeapp/treeprint-form.html` als Zeichnung erstellen lassen.

zur Landkarte für Pauschalreisende, Abschnitt 1.