Bioinformatik: Vergleichende Genetik

Was ist Bioinformatik?

Die Bioinformatik (englisch bioinformatics, auch computational biology) ist eine interdisziplinäre Wissenschaft, die Probleme aus den Lebenswissenschaften mit theoretischen computergestützten Methoden löst. Bioinformatik ist ein weitgefächertes Forschungsgebiet, sowohl bei Problemstellungen als auch den angewandten Methoden. Wesentliche Gebiete sind die Verwaltung und Integration biologischer Daten, die Sequenzanalyse, die Strukturbioinformatik und die Analyse von Daten aus Hochdurchsatzmethoden (~omics). Da Bioinformatik unentbehrlich ist, um Daten in großem Maßstab zu analysieren, bildet sie einen wesentlichen Pfeiler der Systembiologie. Im Folgenden werden einige grundlegende Begriffe erläutert.

Übliche Sequenztypen:

- DNA, RNA, Codons, Proteine

Repräsentation: 1 Letter code DNA/RNA

A	adenosine	C	cytidine	G	guanine
T	thymidine	N	A/G/C/T (any)	U	uridine
K	G/T (keto)	S	G/C (strong)	Y	T/C (pyrimidine)
M	A/C (amino)	W	A/T (weak)	R	G/A (purine)
B	G/T/C	D	G/A/T	H	A/C/T
V	G/C/A	-	gap of indeterminate length		

Repräsentation: 1 Letter code proteins

A	alanine	P	proline
B	aspartate/asparagine	Q	glutamine
C	cystine	R	arginine
D	aspartate	S	serine
E	glutamate	T	threonine
F	phenylalanine	U	selenocysteine
G	glycine	V	valine
H	histidine	W	tryptophan
I	isoleucine	Y	tyrosine
K	lysine	Z	glutamate/glutamine
L	leucine	X	any
M	methionine	*	translation stop
N	asparagine	-	gap of indeterminate length

Repräsentation: Codon Tabelle

Der genetische Standardcode ist in Abbildung 1 gezeigt.

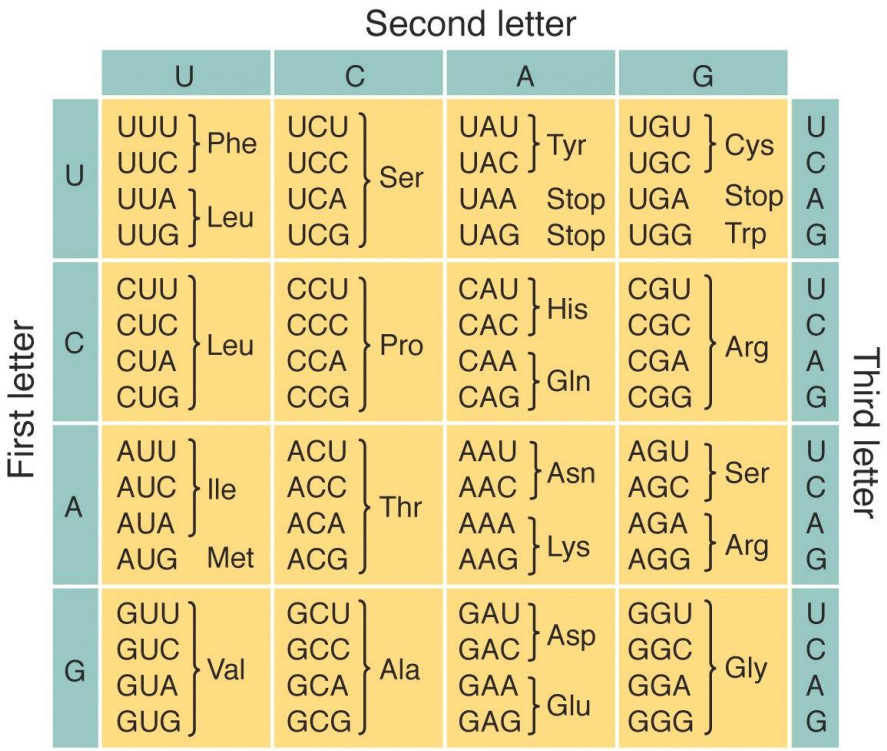

Abbildung 1: Der genetische Standardcode

FASTA Format

Das FASTA-Format ist ein textbasiertes Format zur Darstellung und Speicherung der Primärstruktur von Nukleinsäuren (Nukleinsäuresequenz) und Proteinen (Proteinsequenz) in der Bioinformatik. Die Nukleinbasen bzw. Aminosäuren werden durch einen Ein-Buchstaben-Code dargestellt. Das Format erlaubt es, den Sequenzen einen Namen und Kommentare in einer Kopfzeile (gekenneichnet durch ein >) voranzustellen.

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

Identifiers

Identifiers sind normalerweise einfache Zuriffscodes mit oder ohne Versionsnummer. Ein identifier ist relativ flexibel und meist einfach (zum Beispiel ein einzelnes Wort oder Buchstaben/Zahlenkombination). Leerzeichen (sowohl davor als auch danach bzw. im Identifier selber fuehren meist zu Problemen). Beispiele die zu Problemen fuehren sind:

```
ACCESSION   P01013
AAA68881. 1
gi| 129295
```

Korrekte Identifier dagegen sind

```
ACCESSION_P01013
AAA68881.1
gi|129295
```

Alignment

Sequenzalignment (von lateinisch sequentia, „Aufeinanderfolge“ und englisch alignment, „Abgleich, Anordnung, Ausrichtung“) bezeichnet den methodischen Vergleich zweier oder mehrerer Nukleotid- oder Aminosäuresequenzen in lineare Abfolge. Es wird u.a. in der molekularen Phylogenie verwendet, um die funktionelle oder evolutionäre Verwandtschaft (Homologie) von Nukleotidsequenzen, Aminosäuresequenzen oder Codonsequenzen zu untersuchen. Ein Beispiel für ein multiples Proteinalignment ist Ihnen in Abbildung 2 gezeigt.

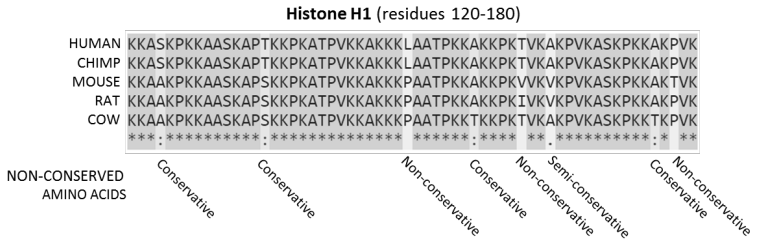

Abbildung 2: Ein Beispiel für ein multiples Sequenzalignment mir einer annotierten Angabe des Konservierungsgrades

Das komplette menschliche Genom

Das Telomere-to-Telomere (T2T)-Konsortium hat einen gewaltigen Schritt in der Humangenomforschung getan. In der Zeitschrift Science beschreibt die Gruppe die erste im Wesentlichen vollständige Sequenz eines menschlichen Genoms, die sogenannte T2T-CHM13-Sequenz. Diese bahnbrechende Ressource, die Regionen des Genoms enthält, die zuvor nicht in einer Referenzsequenz enthalten waren, stellt eine alternative Sequenz zum bestehenden menschlichen Referenzgenom dar. Zusammen mit den ausgefeilten Methoden, die seine Erstellung ermöglichten, eröffnet die Assemblierung einen Weg zur Erzeugung vieler verschiedener menschlicher Genome.

Das neue Genom

In der Ihnen vorliegenden Publikation (Nurk et al., 2022) wird in Tabelle 1 das neue T2T-CHM13 Genom mit dem bisherigen menschlichen Referenzgenom GRCH38 verglichen. Beantworten Sie folgende Fragen mit Richtig/Falsch

(1) Durch T2T-CHM13 sind insbesondere neue protein-codierende Gene entdeckt worden.

(2) Das Verhältnis der Anzahl von Basenpaaren von rDNA Abschnitten zu segmentellen Duplikationen hat sich erhöht.

(3) Insbesondere Satelliten DNA konnte im neuen Genom besser aufgelöst werden.

(4) Das neue Genom hat mehr zusammenhängende Abschnitte (contigs) und mehr menschenspezifische ("exclusive") Gene.

(5) Das Y-Chromosom ist nicht im neuen Genom berücksichtigt.

(6) Das neue Genom ist weitesgehend homozygot.

Wiederholende Sequenzabschnitte / Repeatelemente

Insbesondere sich wiederholende Sequenzabschnitte spielen

(7) In der Datei **RepeatExamples.fa** sind Ihnen verschiedene Repeatelemente gegeben: LTR1, (CCTA)n, 5S, 6kbHsap, AluYd2. Ordnen Sie diese 5 Beispiele ein:

SINE: …
LINE: …
Simple Repeat: …
Low complexity Region: …
rDNA: …
Satellite: …

Das Alter von Repeatelementen kann mithilfe der Kimuradistanz abgeschätzt werden, wobei eine höhere Kimura Distanz ältere Repeatelemente darstellt. In den Abbildungen 3 sind die Häufigkeiten von bestimmten Repeatelementen für verschiedene Altersklassen gegeben (*Repeatlandscapes*). Ein Beispiel (Abbildung 4) einer anderen Spezies zeigt Ihnen, dass Repeatlandscapes sehr unterschiedlich aussehen können.

(8) Welche Repeatelemente haben eine kürzliche Expansion in der Evolution zum menschlichen Genom durchgeführt? …

Unter https://www.repeatmasker.org/genomicDatasets/RMGenomicDatasets.html finden Sie weitere

Repeatlandscapes: Schimpanse (https://www.repeatmasker.org/species/panTro.html), Kalong (https://www.repeatmasker.org/species/pteVam.html), Krokodil (https://www.repeatmasker.org/species/croPor.html), Zebrafisch (https://www.repeatmasker.org/species/danRer.html), Zebrafink (https://www.repeatmasker.org/species/taeGut.html)

(9) Welches der 5 Genome ist vermutlich das kleinste? …

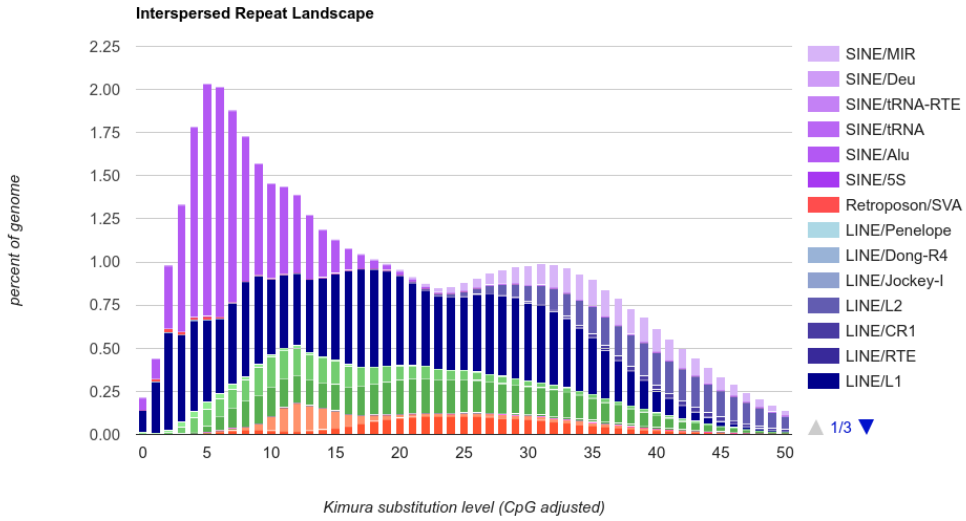

Abbildung 3: Wiederholungselemente beim Menschen

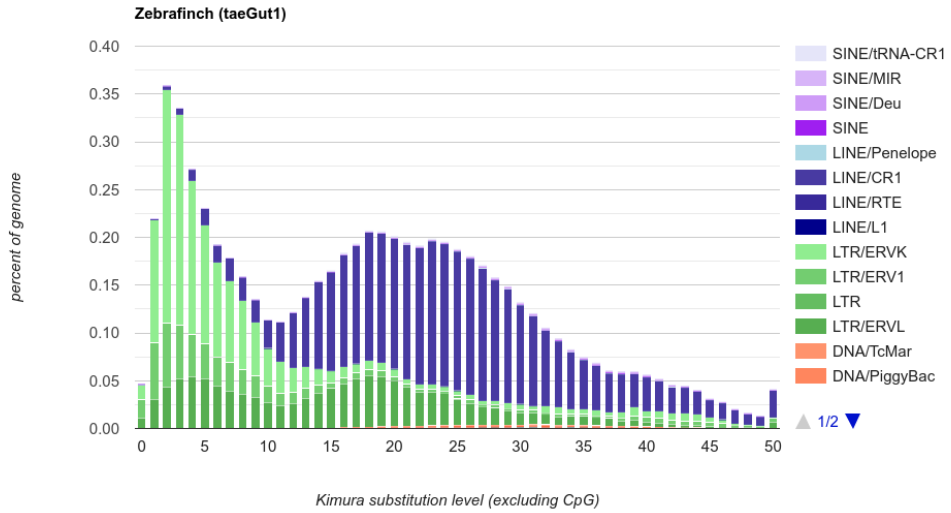

Abbildung 4: Wiederholungselemente beim Zebrafinken

Alu Elemente

Aluelemente sind primatenspezifische Wiederholungen und machen 11 % des menschlichen Genoms aus. Sie haben einen weitreichenden Einfluss auf die Genexpression.

In der Datei **AluSeq.fa** sind Ihnen Sequenzen von Alu Elementen gegeben. Führen Sie zunächst ein Alignment durch mithilfe von MUSCLE (https://www.ebi.ac.uk/Tools/msa/muscle/). Erstellen Sie dabei auch einen Baum. In Abbildung 5 finden Sie einen Stammbaum der Primaten ALU Elemente.

Tragen Sie bitte die Subfamilien (Y, S, J, K, L) in den entsprechenden Baum basierend auf Ihren Alignment Daten.

(10) a: …

(11) b: …

(12) c: …

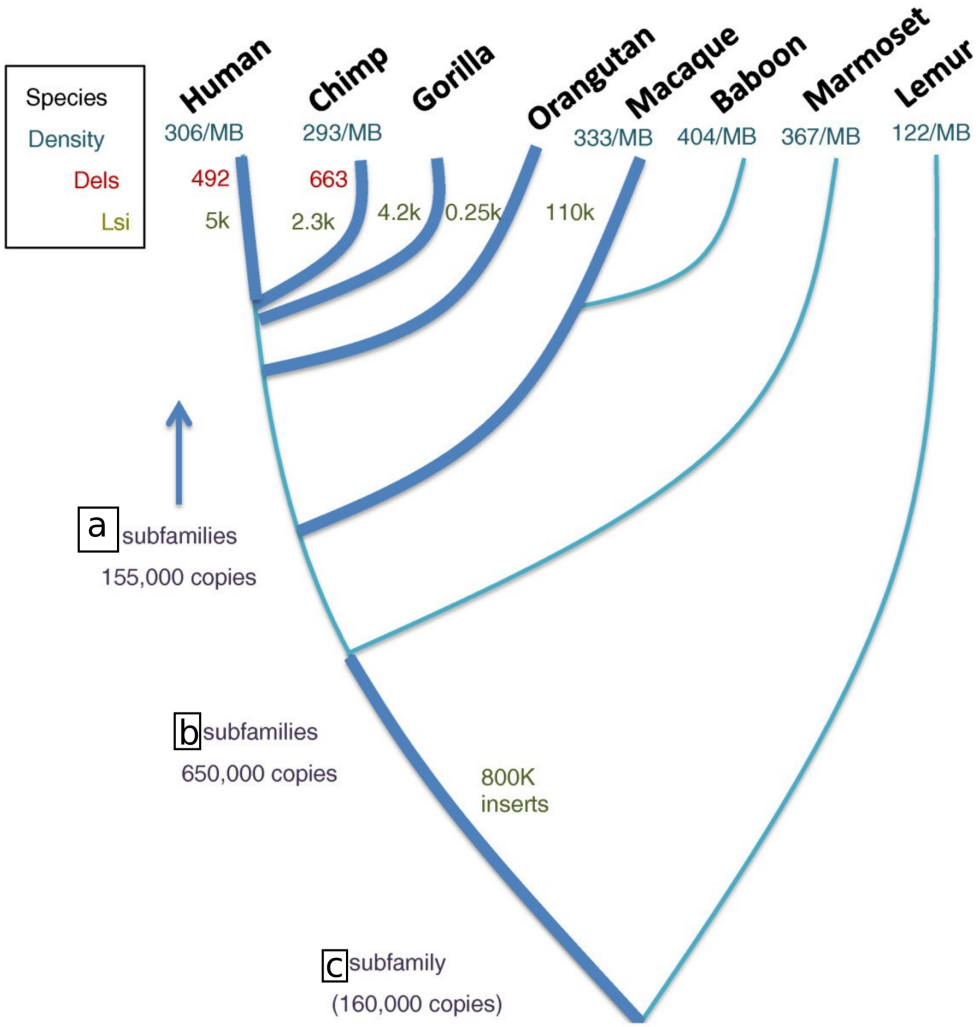

Abbildung 5: Stammbaum der Alu repeates in Primaten

Die neuen Chromosomen

Ziel einer Sequenzierung ist es, die Erbinformation, die in den Chromosomen enthalten ist, in Form einer langen Textzeile zu erhalten. Die verwendeten Sequenziermaschinen sind jedoch nicht in der Lage, das gesamte Genom in einem Schritt zu lesen. Aus diesem Grund muss das Genom in kleine Stücke zerteilt werden, deren Buchstabenfolge gelesen werden kann. Um diese Einzelteile später wieder zusammensetzen zu können, müssen die einzelnen Bruchstücke überlappen. Dies kann man dadurch erreichen, indem man viele Kopien eines DNA-Einzelstranges erzeugt und jede Kopie zufällig zerstückelt (zum Beispiel mit Ultraschall, hoher Druck). Die vielen Einzelabschnitte werden anschließend von Maschinen gelesen.

Die Gesamtabfolge der Bausteine wird mit Hilfe der Überlappungsinformationen gewonnen: Schrittweise werden Einzelstücke gesucht, deren Enden überlappen. Diese überlappenden Abschnitte werden in der entsprechenden Reihenfolge zusammengefügt. Diesen Vorgang nennt man Assemblierung.

Abbildung 1 aus Nurk et al, 2022 illustriert genomische Eigenschaften der des neuen Genom-Assemblies. Welches sind relativ zur Chromosomenlänge die am umfangreichsten erneuerten Chromosomen? (Abbildung 2A/B)

(13) Top1:

(14) Top2:

(15) Top3:

(16) Welche Chromosomen sind acrozentrisch? …

Assemblierungsgraphen

Um Assemblierungsgraphen aus Abbildung 2 von Nurk et al. besser zu verstehen benutzen wir das Programm bandage (https://github.com/rrwick/Bandage/releases/download/v0.8.1/Bandage_Windows_v0_8_1.zip). Bitte laden Sie das Programm herunter, entpacken Sie es und starten Sie es. Anschließend sind Ihnen einen Datensatz zur Verfügung gestellt (**Bacterium_LastGraph**).

Der erste Datensatz enthält eines Assemblierungsvisualisierung eines Bakterium. Nutzen Sie dazu die eingebette BLAST-funktion (eine Tool um Online Datenbankabgleiche durchzuführen).

(17) Um welche bakterielle Spezies handelt es sich? …

Sie sehen 3 größere zusammenhänge Graphen. Welche genomischen Einheiten entsprechen diese? Nutzen Sie dazu die eingebette BLAST-funktion.

(18) Größter Graph: …

(19) Zweitgrößter Graph: …

(20) Drittgrößter Graph: …

(21) In Abbildung 2 (Nurk et al) ist das Genom durch Graphen dargestellt. Die meisten Chromosomen visualieren sich sehr gut als einzelne Abschnitte. Welche Chromosomen teilen sich jedoch wiederholende Abschnitte und können daher im Graphen nicht einzeln aufgelöst werden? Welches Repeatelement ist dafür verantwortlich?

- Chromosomen: …
- Repeatelement: …

CpG content in Repeats

Im neuen T2T Genom konnten auch epigenetische Muster, insbesondere DNA Methylierungsstellen gefunden werden.

(22) Wie viele zusätzliche CpG Dinukleotide konnten im T2T Genom gefunden werden? …

Insbesondere sind bei der Regulation CpG Inseln von Interesse. Ein einfaches Tool zur Berechnung von CpG Insel-Vorkommen steht Ihnen hier zur Verfügung: https://www.bioinformatics.nl/cgi-bin/emboss/cpgplot

(23) Welche Formel bzw. Definition benutzt dieser Rechner als Standardeinstellung

(24) Testen Sie die Repeatelemente LTR1, 5S, 6kbHsap, AluYd2 aus **RepeatExamples.fa** darauf, ob diese CpG Inseln bilden.

(25) Geben Sie eine kodierenden Sequenz an, die eine CpG Insel darstellt, jedoch kein Arginin, Serin, Prolin, Threonin oder Alanin enthält. Nutzen Sie den genetischen Code aus der Einleitung.

(26) Im Folgenden sind Ihnen kurze Wiederholungssequenzen gegeben. Markieren Sie alle Repeats die CpG Inseln formen könnten:

```
>(AC)n  Simple Repeat   Eukaryota\
acacacacacacacacacacacacacacacacacacacacacacacacacacacacacacacacacaca\
>(AG)n  Simple Repeat   Eukaryota\
agagagagagagagagagagagagagagagagagagagagagagagagagagagagagagagagaga\
>(AT)n  Simple Repeat   Eukaryota\
atatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatatata\
>(C)n   Simple Repeat   Eukaryota\
ccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccc\
>(CAA)n Simple Repeat   Eukaryota\
caacaacaacaacaacaacaacaacaacaacaacaacaacaacaacaacaacaacaacaacaacaacaa\
>(CAAA)n    Simple Repeat   Eukaryota\
caaacaaacaaacaaacaaacaaacaaacaaacaaacaaacaaacaaacaaacaaacaaacaaacaaac\
>(CAAAA)n   Simple Repeat   Eukaryota\
caaaacaaaacaaaacaaaacaaaacaaaacaaaacaaaacaaaacaaaacaaaacaaaacaaaacaaa\
>(CAAAAA)n  Simple Repeat   Eukaryota\
caaaaacaaaaacaaaaacaaaaacaaaaacaaaaacaaaaacaaaaacaaaaacaaaaacaaaaacaa\
>(CCA)n Simple Repeat   Eukaryota\
ccaccaccaccaccaccaccaccaccaccaccaccaccaccaccaccaccaccaccaccaccaccacca\
>(CCCCA)n   Simple Repeat   Eukaryota\
ccccaccccaccccaccccaccccaccccaccccaccccaccccaccccaccccaccccaccccaccc\
>(CCCCAA)n  Simple Repeat   Eukaryota\
ccccaaccccaaccccaaccccaaccccaaccccaaccccaaccccaaccccaaccccaaccccaaccc\
>(CCCCCA)n  Simple Repeat   Eukaryota\
ccccaccccccaccccccaccccccaccccccaccccccaccccccaccccccaccccccaccccccc\
>(CCCGAA)n  Simple Repeat   Eukaryota\
cccgaacccgaacccgaacccgaacccgaacccgaacccgaacccgaacccgaacccgaacccgaaccc\
>(CCCTAA)n  Simple Repeat   Eukaryota\
ccctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaaccc\
>(CCCTCA)n  Simple Repeat   Eukaryota\
ccctcaccctcaccctcaccctcaccctcaccctcaccctcaccctcaccctcaccctcaccctcaccc\
>(CCTA)n    Simple Repeat   Eukaryota\
cctacctacctacctacctacctacctacctacctacctacctacctacctacctacctacctacctac\
>(CG)n  Simple Repeat   Eukaryota\
cgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgc\
>(CGAA)n    Simple Repeat   Eukaryota\
cgaacgaacgaacgaacgaacgaacgaacgaacgaacgaacgaacgaacgaacgaacgaacgaacgaac\
```

```
>(CGGA)n    Simple Repeat   Eukaryota\
cggacggacggacggacggacggacggacggacggacggacggacggacggacggacggacggac\
>(CTA)n Simple Repeat   Eukaryota\
ctactactactactactactactactactactactactactactactactactactactactacta\
>(CTCCA)n   Simple Repeat   Eukaryota\
ctccactccactccactccactccactccactccactccactccactccactccactccactcc\
>(GAA)n Simple Repeat   Eukaryota\
gaagaagaagaagaagaagaagaagaagaagaagaagaagaagaagaagaagaagaagaagaagaa\
>(GAAA)n    Simple Repeat   Eukaryota\
gaaagaaagaaagaaagaaagaaagaaagaaagaaagaaagaaagaaagaaagaaagaaagaaagaaag\
>(GAAAA)n   Simple Repeat   Eukaryota\
gaaaagaaaagaaaagaaaagaaaagaaaagaaaagaaaagaaaagaaaagaaaagaaaagaaaagaaa\
>(GAAAAA)n  Simple Repeat   Eukaryota\
gaaaaagaaaaagaaaaagaaaaagaaaaagaaaaagaaaaagaaaaagaaaaagaaaaagaaaaagaa\
>(GACA)n    Simple Repeat   Eukaryota\
gacagacagacagacagacagacagacagacagacagacagacagacagacagacagacagacagacag\
>(GAGACA)n  Simple Repeat   Eukaryota\
gagacagagacagagacagagacagagacagagacagagacagagacagagacagagacagagacagag\
>(GCA)n Simple Repeat   Eukaryota\
gcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagca\
>(GCC)n Simple Repeat   Eukaryota\
gccgccgccgccgccgccgccgccgccgccgccgccgccgccgccgccgccgccgccgccgccgcc\
>(GCCA)n    Simple Repeat   Eukaryota\
gccagccagccagccagccagccagccagccagccagccagccagccagccagccagccagccagccag\
>(GCCC)n    Simple Repeat   Eukaryota\
gcccgcccgcccgcccgcccgcccgcccgcccgcccgcccgcccgcccgcccgcccgcccgcccgcccg\
>(GCCCA)n   Simple Repeat   Eukaryota\
gcccagcccagcccagcccagcccagcccagcccagcccagcccagcccagcccagcccagcccagccc\
>(GCCCC)n   Simple Repeat   Eukaryota\
gccccgccccgccccgccccgccccgccccgccccgccccgccccgccccgccccgccccgccccgccc\
>(GCCCCA)n  Simple Repeat   Eukaryota\
gccccagccccagccccagccccagccccagccccagccccagccccagccccagccccagccccagcc\
>(GCCCCC)n  Simple Repeat   Eukaryota\
gcccccgcccccgcccccgcccccgcccccgcccccgcccccgcccccgcccccgcccccgcccccgcc\
>(GCGCA)n   Simple Repeat   Eukaryota\
gcgcagcgcagcgcagcgcagcgcagcgcagcgcagcgcagcgcagcgcagcgcagcgcagcgcagcgc\
>(GCTCA)n   Simple Repeat   Eukaryota\
gctcagctcagctcagctcagctcagctcagctcagctcagctcagctcagctcagctcagctcagctc\
>(GGA)n Simple Repeat   Eukaryota\
ggaggaggaggaggaggaggaggaggaggaggaggaggaggaggaggaggaggaggaggaggaggaga\
>(GGAA)n    Simple Repeat   Eukaryota\
ggaaggaaggaaggaaggaaggaaggaaggaaggaaggaaggaaggaaggaaggaaggaaggaaggaag\
>(GGAGA)n   Simple Repeat   Eukaryota\
ggagaggagaggagaggagaggagaggagaggagaggagaggagaggagaggagaggagaggagaggag\
>(GGAGAA)n  Simple Repeat   Eukaryota\
ggagaaggagaaggagaaggagaaggagaaggagaaggagaaggagaaggagaaggagaaggagaaggaga\
>(GGCA)n    Simple Repeat   Eukaryota\
ggcaggcaggcaggcaggcaggcaggcaggcaggcaggcaggcaggcaggcaggcaggcaggcaggcag\
>(GGCCC)n   Simple Repeat   Eukaryota\
ggcccggcccggcccggcccggcccggcccggcccggcccggcccggcccggcccggcccggcccggcc\
>(GGGA)n    Simple Repeat   Eukaryota\
gggagggagggagggagggagggagggagggagggagggagggagggagggagggagggagggagggag\
>(GGGAGA)n  Simple Repeat   Eukaryota\
gggagagggagagggagagggagagggagagggagagggagagggagagggagagggagagggagaggg\
>(GGGGA)n   Simple Repeat   Eukaryota\
ggggaggggaggggaggggaggggaggggaggggaggggaggggaggggaggggaggggaggggaggggg\
>(GGGGGA)n  Simple Repeat   Eukaryota\
gggggaggggaggggaggggaggggaggggaggggaggggaggggaggggaggggaggggaggggaggg\
```