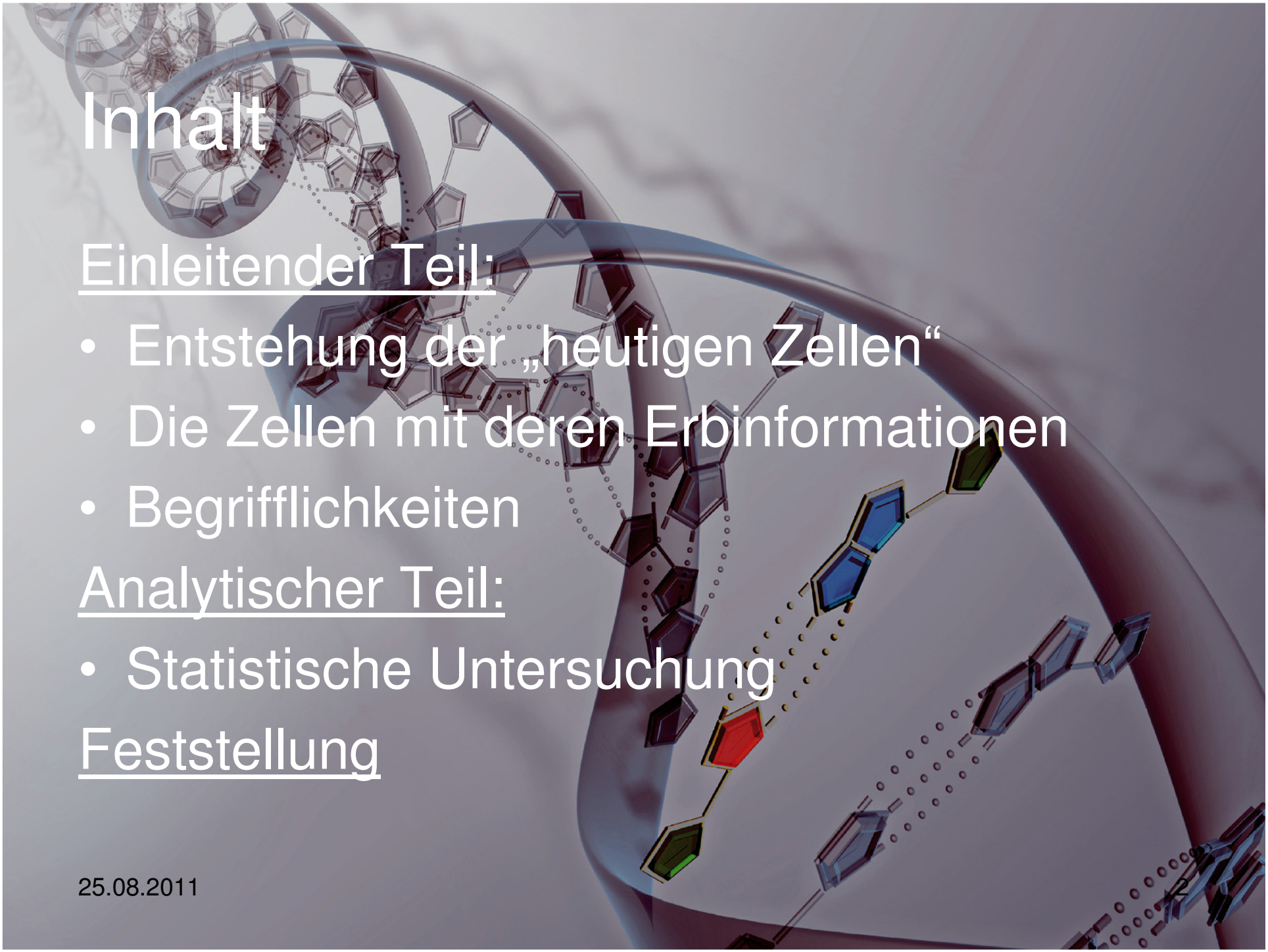
Seminar „*Genetischer Code und Genomgrammatik*“

Thema: „Statistische Betrachtung des menschlichen Genoms“

Dozent: Prof. Dr. rer. nat. Andreas de Vries

Studierender: Cem Kiyak

25.08.2011



Inhalt

Einleitender Teil:

- Entstehung der „heutigen Zellen“
- Die Zellen mit deren Erbinformationen
- Begrifflichkeiten

Analytischer Teil:

- Statistische Untersuchung

Feststellung

Entstehung der „heutigen Zellen“

- Leben auf der Erde etwa 3,5 Milliarden Jahre alt
- drei große Domänen des Lebens:
  Eukaryonten,
  Eubakterien und Archaebakterium

- <u>Annahme:</u>
  Verschmelzung des Eubakteriums sowie Archaebakteriums für die Entstehung des Lebens

Die Zellen mit deren Erbinformationen

- Der Körper besteht aus rund **100.000.000.000.000 Zellen**
- Jede biologische Art, jede Spezies, ist in ihrer Erbinformation einzigartig, sondern sogar jedes Individuum
- **Ausnahme:** Klon

Arbeiten aus der Praxis: Auf der individuellen Einzigartigkeit basieren forensische Analysen

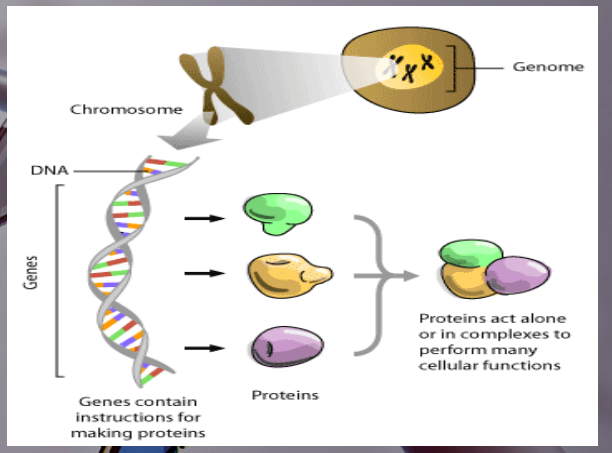
- Die Biologie ist eine Wissenschaft der Grauzonen und Ausnahmen
- Beispiel: **Rote Blutkörperchen**, funktionierende Zellen, aber keine Erbinformation mehr
- Können sich nicht teilen und müssen aus den Stammzellen im Knochenmark gebildet werden
- **Viren** haben Erbinformation, aber sie leben nicht, denn sie brauchen immer lebende Zellen, um Ihre Erbinformation und ihre Hülle zu vermehren.

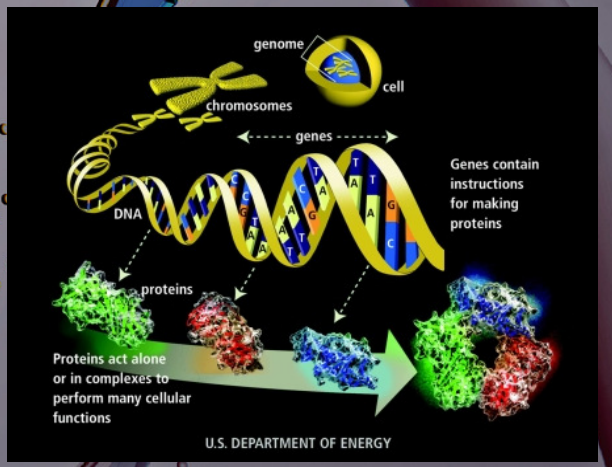
Die Zellen des Immunsystems

- Erbinformation nicht völlig identisch mit anderen Körperzellen,
- Ihre Erbinformation ist etwas umgestaltet (rekombiniert)

Begrifflichkeiten

- Genetik
- Genom und seine Bedeutung
- Gene der Population
- Gene
- DNA





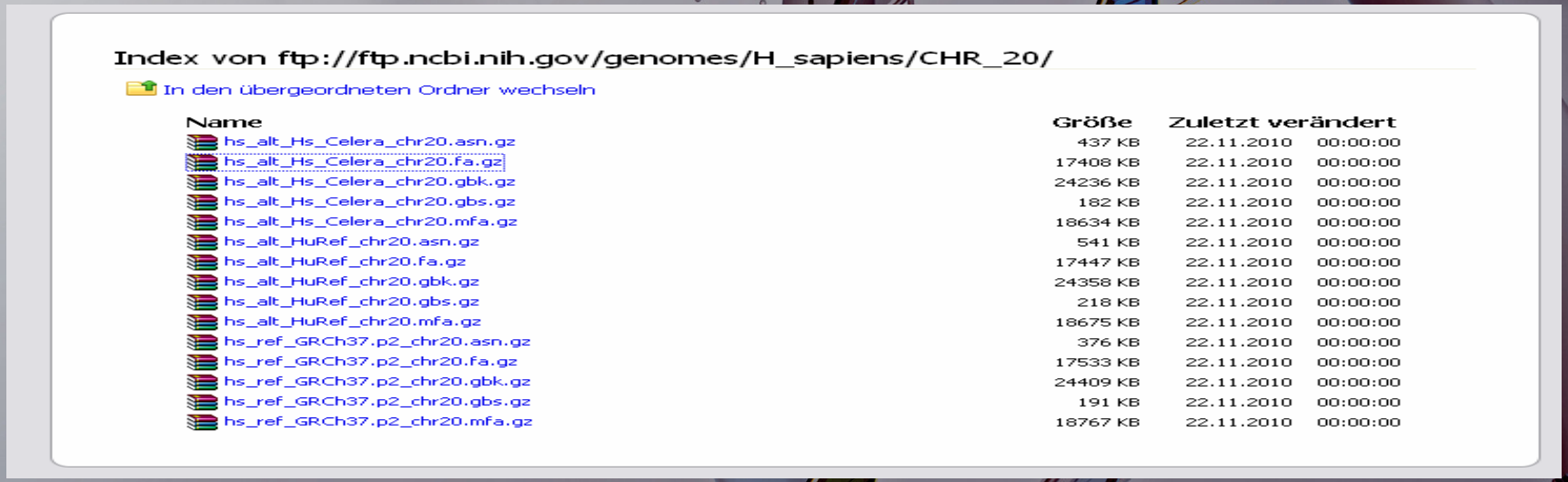
Charakterisierung eines Gens

- Gen eindeutig charakterisiert
- Durch chemische Bausteine :
  vier Nukleotiden Adenosin, Cytidin, Guanosin und Thymidin
- Baukastenprinzip
- Abfolge dieser Nukleotide, sogenannte Nukleotidsequenz ist ein Gen eindeutig beschrieben
- Drei Milliarden Basenpaare in der DNA des Menschen machen unser Genom aus

Vorgehensweise

- NCBI: National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information aus den USA
- Datenbank der Genomsequenzen per FTP

  http://www.ncbi.nlm.nih.gov/



Index von ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/CHR_20/

In den übergeordneten Ordner wechseln

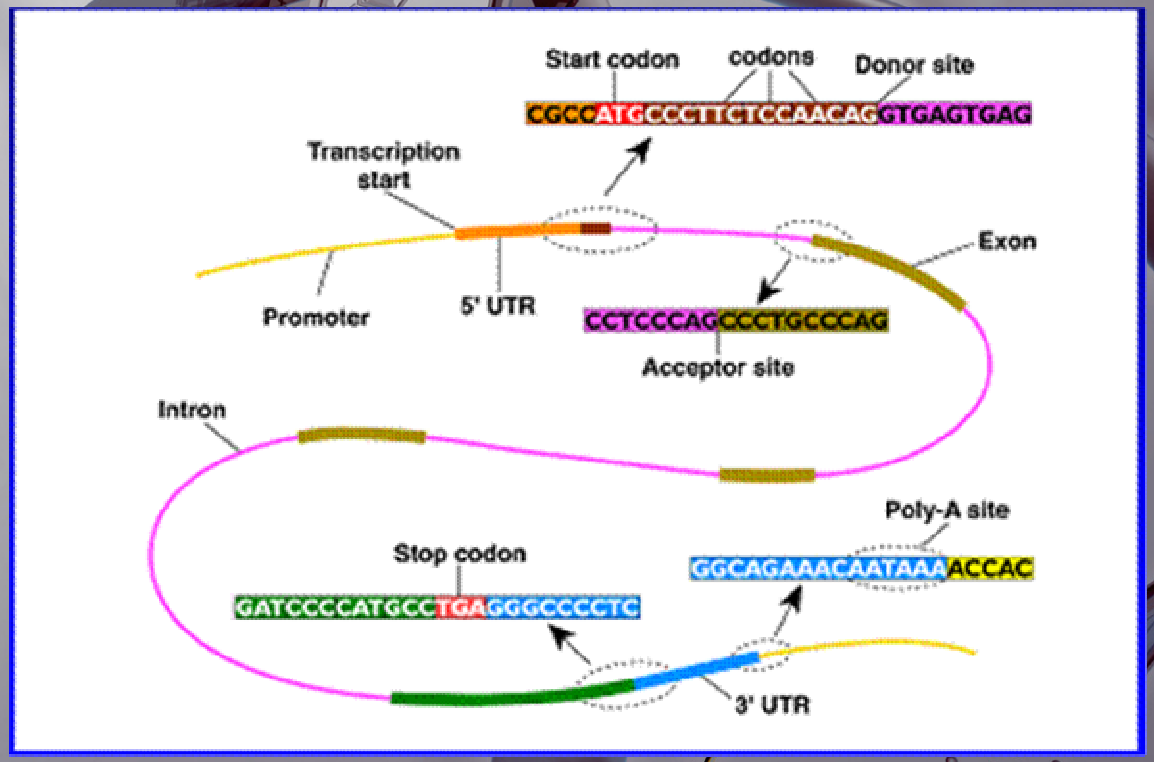
Name	Größe	Zuletzt verändert
hs_alt_Hs_Celera_chr20.asn.gz	437 KB	22.11.2010 00:00:00
hs_alt_Hs_Celera_chr20.fa.gz	17408 KB	22.11.2010 00:00:00
hs_alt_Hs_Celera_chr20.gbk.gz	24236 KB	22.11.2010 00:00:00
hs_alt_Hs_Celera_chr20.gbs.gz	182 KB	22.11.2010 00:00:00
hs_alt_Hs_Celera_chr20.mfa.gz	18634 KB	22.11.2010 00:00:00
hs_alt_HuRef_chr20.asn.gz	541 KB	22.11.2010 00:00:00
hs_alt_HuRef_chr20.fa.gz	17447 KB	22.11.2010 00:00:00
hs_alt_HuRef_chr20.gbk.gz	24358 KB	22.11.2010 00:00:00
hs_alt_HuRef_chr20.gbs.gz	218 KB	22.11.2010 00:00:00
hs_alt_HuRef_chr20.mfa.gz	18675 KB	22.11.2010 00:00:00
hs_ref_GRCh37.p2_chr20.asn.gz	376 KB	22.11.2010 00:00:00
hs_ref_GRCh37.p2_chr20.fa.gz	17533 KB	22.11.2010 00:00:00
hs_ref_GRCh37.p2_chr20.gbk.gz	24409 KB	22.11.2010 00:00:00
hs_ref_GRCh37.p2_chr20.gbs.gz	191 KB	22.11.2010 00:00:00
hs_ref_GRCh37.p2_chr20.mfa.gz	18767 KB	22.11.2010 00:00:00

Java-Tool

- Durch das Java-Tool können nachvollziehbare statistische Auswertungen der FASTA-Dateien bearbeitet werden
- Download unter:

  http://www.math-it.org/java

Aufbau des Stranges



Beispiel 1/2 zur Betrachtung des Chromosoms Nr. 20

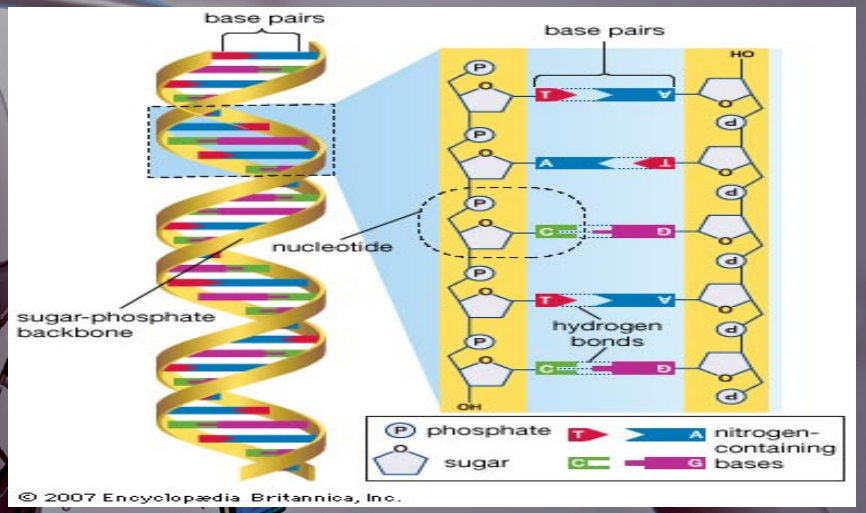
CAAATACAAGGGAGAAATTATGATATTCTCAGATAAACAAAAGCTAACAGAGTTCATGACCATTAAACCT
TGCCTACAAGAAATGCTAAAGGGAGTGCTTCAGGTGGAAATGAAAGGATGCTATATGGCAACTCAAAGTT
GTATGAAGAAATAAAGATATCCAGTTATGTGTACATTTTTTAAAAATCATGCTTCAATATCCCACAAACA
AGCTGTGAGCACCCTGCCCAGGTGACCAGGTAACCAGTATTTTAAAGCTGGTCCCTGCCACAATTCCTCT
TCTTGTCTTCCCCTACTGTGACCTAGTGACATTGAAACACTCTAGTAAAAGTGTTTTTGCTTGTGTCCTC
CACCTTGACTCCCAGTAAAGGCACTTGCCCAGGGATCCCATCTCTCTTGCCTACCCATCTGCTTGATTGA
GCCTGCTCCTTGGGAGCTCCTTCCACATGGCTTCCTGCATGGCCTGCCCTACTTCTGTCTCTAGGAACTG
TGAGTATGCCTCTTCATATGAACTTTTCATGGCCATGTTAGAGTGATCCTCTAAGATTCAACCAGCAGGA
GTCACTCTAACATTTTCTATAGTAATGGCAATGAGGATGGGATCATTCTAAACTCAGGTGGAAAGTTCCT
AGGGAATGCCTTTGACTCATGACTTACTGATTTGCTACTCTGATTGGCTCCACAATTTGAGAAGGCTTTT
TACCCTGATGACTTTCTTGTGCTATGCTATGTACAGCTTTGTGTTGCCTTTTGGAGTGCTGCCATGGCTC
TCCATCCTGAAGCGAGGATCCTGTTCCAATATCAATAATGACATCCACCTTTCTCCAAGAGCACAAGATG
ATATAACATTGGTAACACCTCTTCAAAGAAGATTCATGTGATATGCATATTAATGTCACATAATCCAGCA
CCTGTAAGTTTTTTTTGGGATGCCTGTGGCAACATATTCTTCATTTACAAATTTTACTTACAAATTAAAA
TCAACAGACACTCCTGTTAGTGTCATTAAATGAATTAAACAAAAATATCTCTCCTCTGTAAAGACTTTGG
CTGTGTCTTTCATGCCTCTTGGAGTCTCTAGACCACTCATGATGGCTATATATTTCCTCAGGGCCTTGAT
GCAAAAATAATGCTCTTTATGGCCTACGCTGTACTCCATTAAAACAACTGACTGGTTACCCACTAGAATC
TCCTGATTTGGATTCCAGATTTGAAGATAATACCCTCCTGTTTCAATCATTAACCATAAATTGCCCCACA
GGCCAGGCTCTGCTCATCTGAAATGACACCCATCGGAATTTAAACTTCCACTGCAACACAAAATTAAAAT
TGCTGCCAGGTACAGCTGCACACCTTTCTGTTGACAGTTGGCATTATTACTACTTGGTTTGTTGATGCAA
CAATCCTCAACCAAATACTAGCAAACTGAATTGAACAGCATATTAGAAGAATTATACAACACGAGCAAGT
GATATTTATTTCTGGAATACAAGGATGTTTCCATGCACAAATGACAAACAATGTAATATATCATATTTAT
AGAAGAAAGACAAGAAACACATGATTAGCTCATTTGATACAGAAGACGATTTGACAAATTTGGGGATCTT
TTTATGACAAAAACCATCAATAAAAGGAATATATGAAAACTTCTATAAAATAATAAAAGTCGTAAACACA
CACTTGATTTTCACTTATCGCTGTCTAATATACTATTTTATTTGCCTATTTATCATGTGTATTGATTACT
GTCATTCTCCCCAGTGGAGATAAGCAACAAAAGGGCAACTACTTTTTGTCTTTTATTCAACAATGTTCTT
TGTTAGGTCCTGGCACATAGCAGGTATTCATCAAATATGTGTATAATGTGTGAATTGCTTAGTGGAACAC

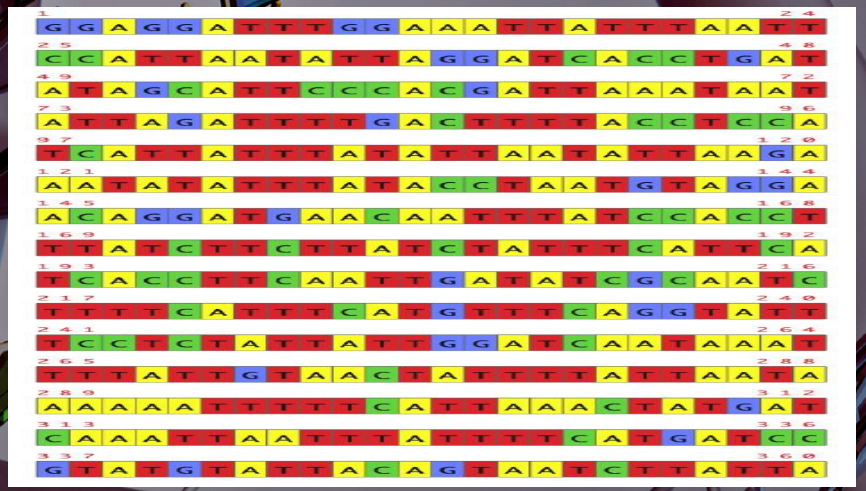
Beispiel 2/2 zur Betrachtung des Chromosoms Nr. 20

```
AGGATCTCTGCATGGGGGCAGGAGTCTCTGCGGGGAGGACTTTTAGGGGGTAACAGGATCTGTGCAGCGT
GCAGGATCTTTGCAGTGGGGCCAGGATCTGTGCAGTGGGTAGGATCTACATTGGGAGTCAGTGTCTGTTC
AAGGGAGTCGGGGGACAGGACCCGTCCAGGGAGGGCAGGATCTGTACAGAGGGACCAAGATCTACGCACG
GGGATAGCATCTATGTGGGGTGGGGGCACGGTCTGTGCCAGGTTTTGGGGGCACTAGATCTTTGCAGGTG
TGGCCAGGATCTGTGCAAGGGGTAGGATCTGTGCAGAGTGGCCAGGATCTGTGCAGAGTGGCCAGGATCT
GTGCAGGGAGGGCAGGATCTGTGCAGAGTGGCCAGGATCTGTGCAGAGTGGCCAGGATCTGTGCAGAGTG
GCCAGGATCTGTGCAGGGAGGCCAGGATCTGTGCAGAGTGGCCANNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

Analyse des menschlichen Genoms nach der Länge der Nukleotidbasen

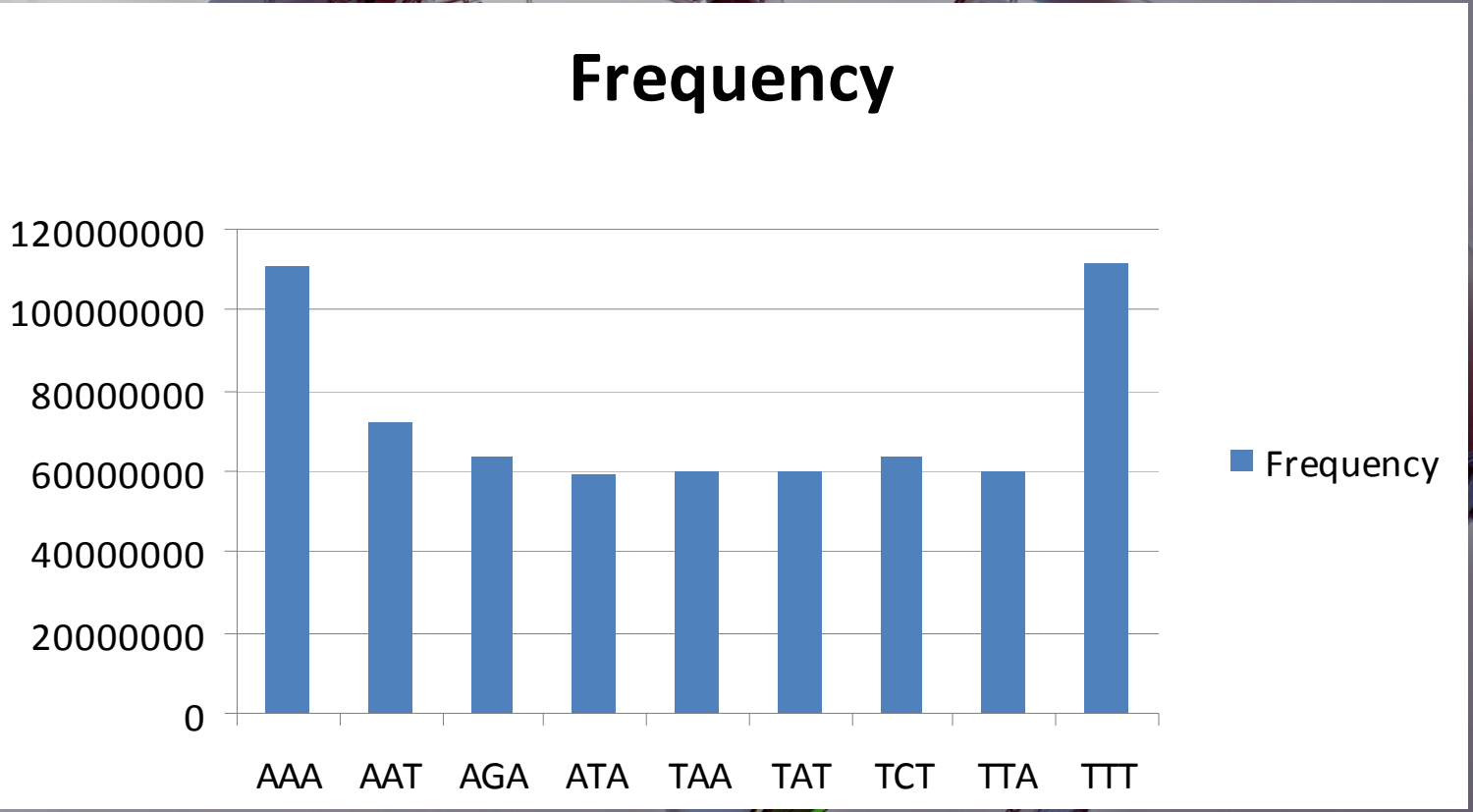
- Nukleotidbasen der Länge 3 (AAA,AAT,AGA,...)
- Nukleotidbasen der Länge 6 (AAAAAA,...)
- Nukleotidbasen der Länge 9 (AAAAAAAAA,...)





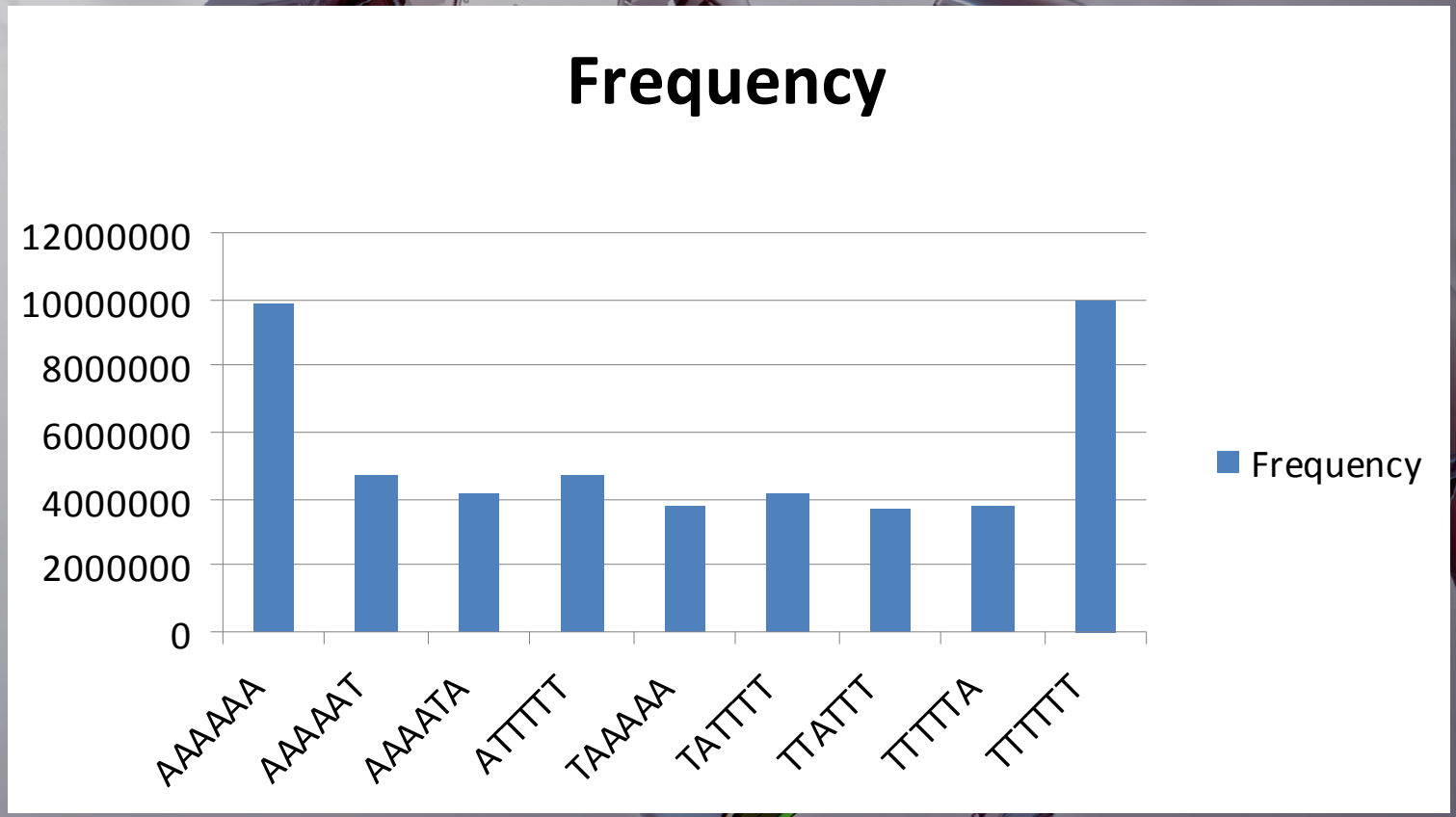
Darstellung der Häufigkeiten

- Nukleotidbasen der Länge 3



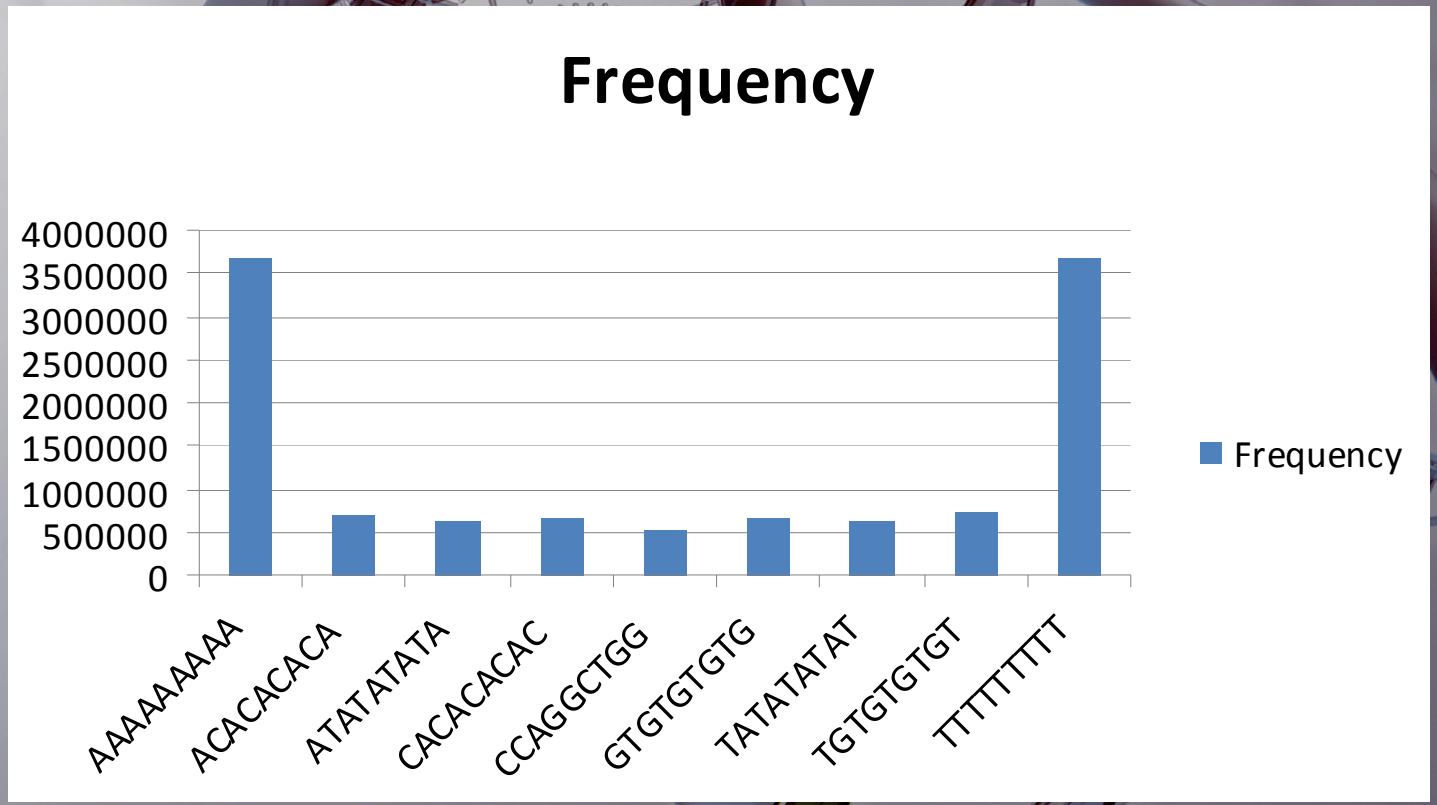
Darstellung der Häufigkeiten

- Darstellung der Häufigkeiten der Länge 6



Darstellung der Häufigkeiten

- Darstellung der Häufigkeiten der Länge 9

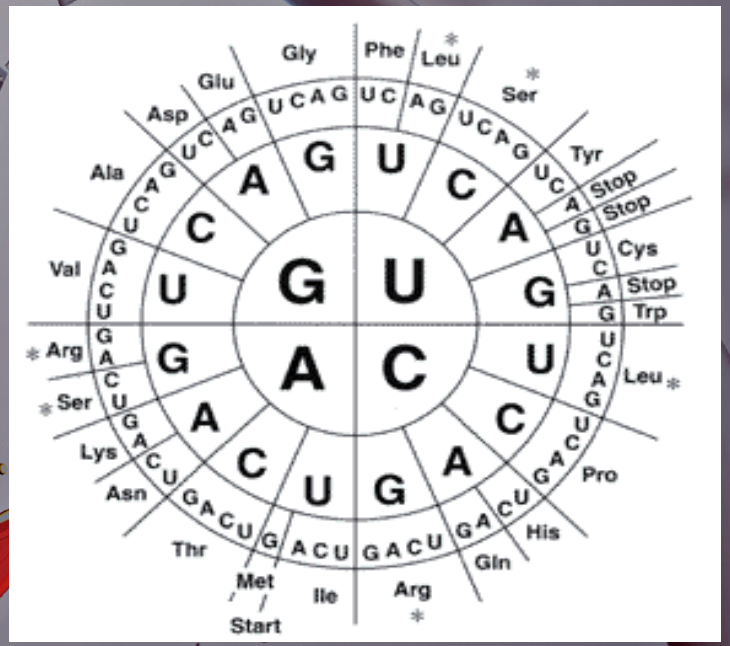


Die nicht genutzten Kombinationen der Länge 9 des menschlichen Genoms

Die 6 häufigsten davon:

- CGTCGATCG 9
- CGACGTACG 7
- CGATACGCG 7
- CGTAACGCG 7
- TACGCGCGA 7
- TCGCGTCGA 7

Codierung der Nucleotide:



Aufsplittung in Condons sowie Aufschlüsselung (Erkenntnis)

Aufsplittung des Genoms in seine Basen (3 davon jeweils = Codons bzw. Triplets) und auffinden seiner Aminosäuren:

• 1) CGT = Arg	CGA = Arg	TCG = Ser
• 2) CGA = Arg	CGT = Arg	ACG = Thr
• 3) CGA = Arg	TAC = Tyr	GCG = Ala
• 4) CGT = Arg	AAC = Asn	GCG = Ala
• 5) TAC = Tyr	GCG = Ala	CGA = Arg
• 6) TCG = Ser	CGT = Arg	CGA = Arg

Nicht vorhandene Kombinationen der Länge 12 (Erkenntnis)

	A	B	C	D	E	F	G	H	I
1	12-letter Words omitted in all human chromosomes								
28									
29	AAAAACGCGCGA	AAATCGGCGACG	AACCGGATGACG	AACGAGTGTTCG	AACGCGGCGCAT	AACGGTCAGTCG	AACTACGTCCGC	AAGCGCTAGCGC	AATAAGTTCGCG
30	AAAAACGCGCTA	AAATCGGCGCGC	AACCGGCACGTA	AACGAGTTCGAT	AACGCGGCGCTA	AACGGTCATGCG	AACTACGTCGCG	AAGCGCTATCCG	AATAATCCGCCG
31	AAAAATCGCGAC	AAATCGGCGCTT	AACCGGCCCGCT	AACGAGTTTACG	AACGCGGCGCTT	AACGGTCATTCG	AACTACGTCGGG	AAGCGCTATGCG	AATAATCCGCGC
32	AAAAATCGCGCG	AAATCGGCGTAA	AACCGGCCCGTA	AACGATAACCGC	AACGCGGCGGAT	AACGGTCCACGA	AACTACTCGTCG	AAGCGCTCGACG	AATAATCGACGG
33	AAAACCGACGCG	AAATCGGCGTCG	AACCGGCCGAAT	AACGATAATCCG	AACGCGGCGTAA	AACGGTCCCGAC	AACTACTGCGCG	AAGCGCTTACGC	AATAATCGCCCG
34	AAAACCGCGCTA	AAATCGGCTCGT	AACCGGCCGCGT	AACGATACACGG	AACGCGGCGTAC	AACGGTCCGACC	AACTAGACGCCG	AAGCGCTTATCG	AATAATCGTCGG
35	AAAACCGGTTCG	AAATCGGGCGAT	AACCGGCGAACG	AACGATACATCG	AACGCGGCTACG	AACGGTCCGACG	AACTAGACGCGT	AAGCGGACAACG	AATACACCGCGA
36	AAAACCGTCGCG	AAATCGGGTACG	AACCGGCGAATT	AACGATACCGCC	AACGCGGCTCGA	AACGGTCCGCCG	AACTAGCCCGCG	AAGCGGACCGGT	AATACACCGGCG
37	AAAACCGTCGCT	AAATCGGGTCGA	AACCGGCGACAT	AACGATACCGCG	AACGCGGCTTCG	AACGGTCCGGTA	AACTAGCCGCGA	AAGCGGACGATA	AATACACCGTCG
38	AAAACGACCGGT	AAATCGGTACGA	AACCGGCGATTA	AACGATACCGGT	AACGCGGGTAGT	AACGGTCCGGTT	AACTAGCCGTCG	AAGCGGACGTAC	AATACACGGCGT
39	AAAACGACGCGC	AAATCGGTCGCC	AACCGGCGCAAT	AACGATACGACC	AACGCGGGTATA	AACGGTCCGTAA	AACTAGCGCCGT	AAGCGGATACGT	AATACCACGTCG
40	AAAACGATCGGG	AAATCGGTTGCG	AACCGGCGCATT	AACGATACGACT	AACGCGGGTCTA	AACGGTCCGTCA	AACTAGCGCGAA	AAGCGGATTCGC	AATACCCGATCG
41	AAAACGCCGATC	AAATCGTAACCG	AACCGGCGCCTA	AACGATACGAGC	AACGCGGGTTAC	AACGGTCCGTCC	AACTAGCGCGCT	AAGCGGATTCGG	AATACCCGCCGA
42	AAAACGCCGTAC	AAATCGTACCCG	AACCGGCGCGCA	AACGATACGCCG	AACGCGGTACGA	AACGGTCCGTCG	AACTAGCGCGTT	AAGCGGCCCGTA	AATACCCGCGAT
43	AAAACGCGACCG	AAATCGTACGGC	AACCGGCGCGGT	AACGATACGCGA	AACGCGGTACGG	AACGGTCCTATC	AACTAGCGCTCG	AAGCGGCGATCT	AATACCCGCGCC
44	AAAACGCGAGCG	AAATCGTCGACG	AACCGGCGCTAA	AACGATACGCGC	AACGCGGTACGT	AACGGTCGAACC	AACTAGCGTACG	AAGCGGCGCCTA	AATACCCGTCGA
45	AAAACGCGATCG	AAATCGTCGCAA	AACCGGCGGCGT	AACGATACGCGG	AACGCGGTACTC	AACGGTCGAATC	AACTAGCGTCGA	AAGCGGCGTACG	AATACCCGTCGG
46	AAAACGCGCGAT	AAATCGTCGCAC	AACCGGCGGTTA	AACGATACGGCG	AACGCGGTAGAA	AACGGTCGACAT	AACTAGCGTGCG	AAGCGGCGTATA	AATACCCTCGCG
47	AAAACGCGCGTA	AAATCGTCGCGC	AACCGGCGTACG	AACGATACGGCT	AACGCGGTATCA	AACGGTCGACCG	AACTAGTACGGG	AAGCGGCTAACG	AATACCGACCGA
48	AAAACGCGTCGG	AAATCGTCGCGT	AACCGGCGTAGT	AACGATACGTTC	AACGCGGTATTC	AACGGTCGACGA	AACTAGTCGCGA	AAGCGGCTATCG	AATACCGACCGG
49	AAAACGCGTTCG	AAATCGTCGCTG	AACCGGCGTCAA	AACGATAGCCGT	AACGCGGTCAAC	AACGGTCGACGC	AACTAGTCGCGC	AAGCGGGTATCG	AATACCGACGAC
50	AAAACGCTTCGC	AAATCGTCGGGG	AACCGGCGTCGT	AACGATAGCGAC	AACGCGGTCAAT	AACGGTCGACTC	AACTAGTCGCGT	AAGCGGGTCCGG	AATACCGACGCG
51	AAAACGGCGCCG	AAATCGTCGGTT	AACCGGCGTGTA	AACGATAGCGCC	AACGCGGTCCAA	AACGGTCGACTG	AACTAGTCGGCG	AAGCGGGTCGAA	AATACCGATCGT
52	AAAACGCGCGGA	AAATCGTCGCGC	AACCGGCGTTAA	AACGATAGCGGC	AACGCGGTCGGA	AACGGTCGACGC	AACTAGTCGCGC	AAGCGGTAATCG	AATACCGCGCGA

Anzahl: 128 000 Kombinationen nicht vorhanden

Was wollen wir bewirken?

- Defekte Gene „reparieren“
- Änderung der Gene um neues oder anderes Leben zu erzeugen
- Krankheiten besser verstehen und evt. Lösungen dazu zu finden

Feststellung

- Intensive weitere Analysen von großer Bedeutung, denn hier ist der Weg das Ziel!
- Verstehen der Codon-Darstellung die Aufschluss auf eine verständliche Sprache ergeben sollen
- Die Chancen könnten gut stehen durch statistische Berechnungen mit durchaus umfangreicheren Untersuchungen und leistungsstarken Rechnern und Geduld ein Ergebnis zu erhalten
- Sind Grammatikalische Strukturen zu erkennen?


Vielen Dank für Ihre Aufmerksamkeit!