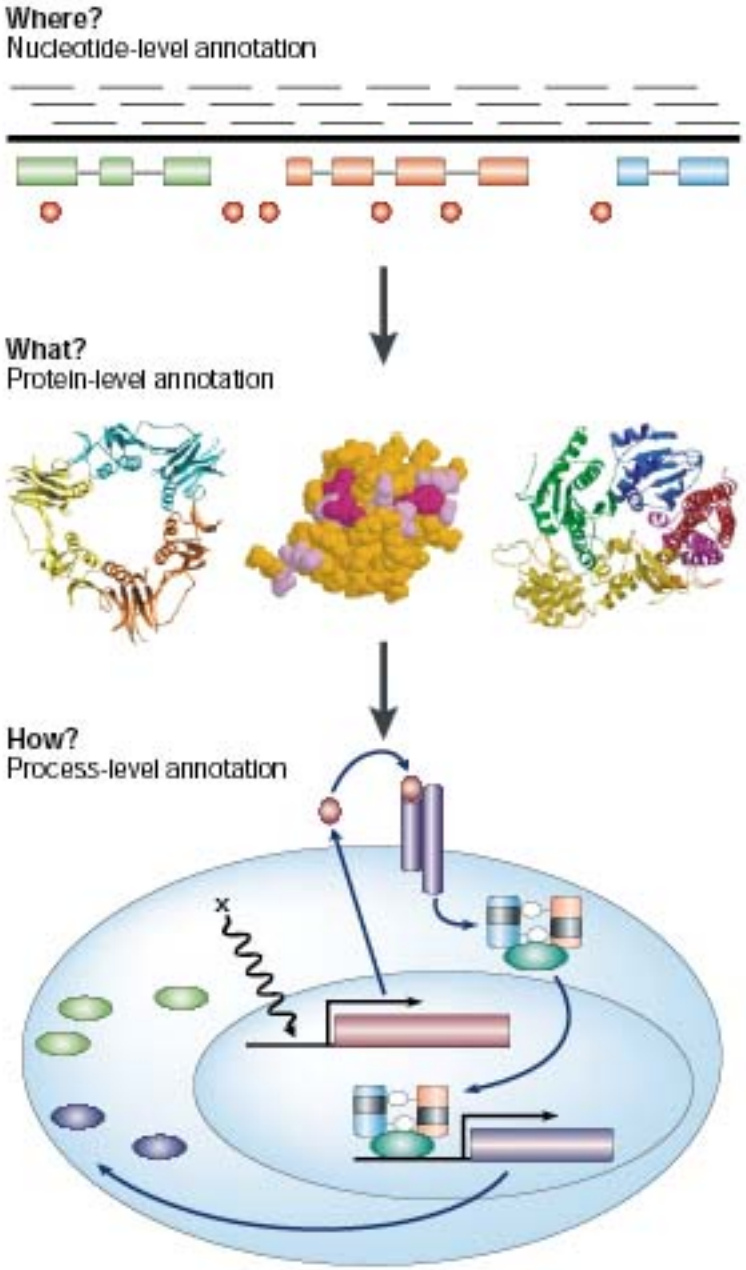
WS2018/2019

„Genomforschung und Sequenzanalyse

- Einführung in Methoden der Bioinformatik- “

Thomas Hankeln

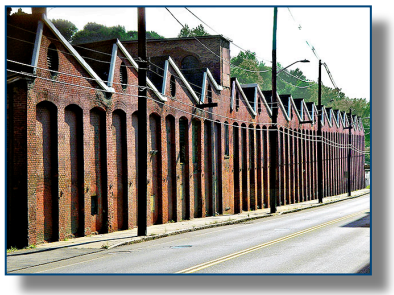
Genvorhersage & Genom-Annotation



Ebenen der Annotation

- Genstruktur (Exons/Introns, UTR ‘s, Promoter, andere regulator. Elemente)
- Orthologe and Paraloge
- Gen-Varianten (SNPs etc.)
- Transkripte (Gewebe, Zelltypen?)
- Protein-Domänen und –motive,
- Kristallstruktur, Faltung
- RNA/Protein: Funktion, intrazell. Lokalisation, Aktivitätsmuster
- interagierende Proteine
- Stoffwechselwege, metabol. Produkte
- Krankheitsrelevanz
- ……

Annotationsstrategien



- „Factory“ > automatisch
- „museum“ > durch Kuratoren
- „party“/ jamboree

- „TPA“ = 3rd party annotation

„GeneOntology“:
Ein gemeinsames Vokabular

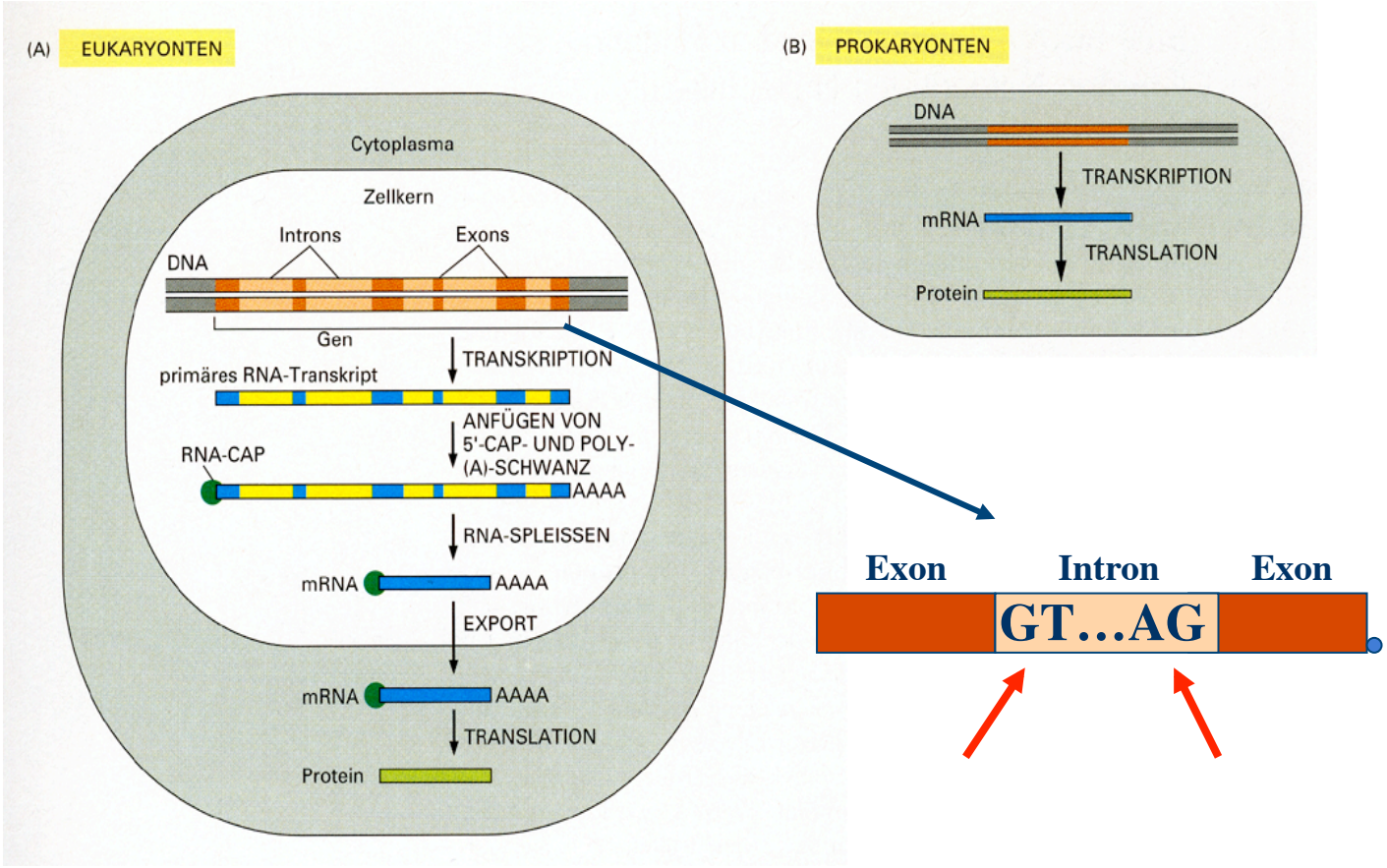
	Bsp: Cytochrome C
„Molecular function“ 7393	electron transporter
„Biological process“ 10370	oxidative phosphorylation
„Cellular component“ 1681	mitochondrial inner membrane

www.geneontology.org

Wo steckt denn nun die genetische Information?

```
  1 ccgaacgctt atagagagct atagagtgaa agctgagaag aaccaaaacg gagcataaac
 61 atgaacagcg atgaggtgca actgatcaag aagacctggg aaatccccgt ggcaacacca
121 acagattctg gagcggcgat actgacgcag tttttcaacc gctttccgtc caacttggag
181 aagttcccct tccgcgatgt tcctttggag gagctaagtg tgagttgtac cttacacata
241 ggtcttcaat taactcaaga ttaacttgat ctgttttctt tcagggaaat gctcgcttcc
301 gagcacatgc cggcagaatc ataagggtct ttgacgagtc catccaggtc ctgggccagg
361 atggcgatct ggagaagctg gacgagatct ggaccaaaat tgccgttagt cacattccgc
421 ggaccgtttc caaggagtct tacaacgtaa gttgaacact gcagtcgagc tctcgacttt
481 gagatacctg ttggtcagat agtggaagtt gaaagctata tgacatttaa aaattcaatt
541 gcatttaaaa catcatttta tttttttag caactgaaag gagttatcct ggatgtgctg
601 acagctgcct gcagtctgga cgagagtcaa gcggccacgt gggccaagct ggtggaccat
661 gtctacgcaa tcatcttcaa ggcgatcgac gacgacggca acgccaagta gatgaggcag
721 ctggaggtgg agatgcaacc gaatccgcgg a
```

Bei Eukaryoten-Genomen ist Generkennung besonders schwierig



Die („vereinfachte“) Aufgabe:

- gegeben sind uncharakterisierte Genom-DNA-Sequenzen

- **FINDE...**

  Protein-kodierende Regionen
  Exon/Intron-Grenzen
  mögliche genregulatorische Abschnitte

Mache daraus ein Modell für die Struktur des Gens!

Warum „vereinfacht“?

- nicht alle Gene werden in Proteine übersetzt!
  (RNA-Gene!!)

- auch nicht alle Genregionen proteinkodierender Gene werden in Proteine übersetzt
  (5‘ und 3‘-untranslatierte Exons)

- Gene werden alternativ gespleißt.
  Die ALT-mRNAs können unterschiedliche Proteine kodieren.

- RNA editing, translational readthrough, ribosome frameshift

Die drei Wege zum Gen

1. Datenbanksuchen/Alignments *extrinsisch*

   „es gibt bereits passende Sequenzen in den Datenbanken“

2. „*ab initio*“-Genvorhersage *intrinsisch*

   „Signale“ und „Inhalt“ in der DNA zeigen: hier ist ein Gen“

3. Vergleich von verwandten Genomen
   („comparative genomics“) *extrinsisch*

   „hier ist eine evolutionär konservierte Region“

Der direkte Beweis:

Gene finden durch DB-Suchen

Ein vermutliches Gen liegt in einem Genom-DNA-Abschnitt vor, wenn...

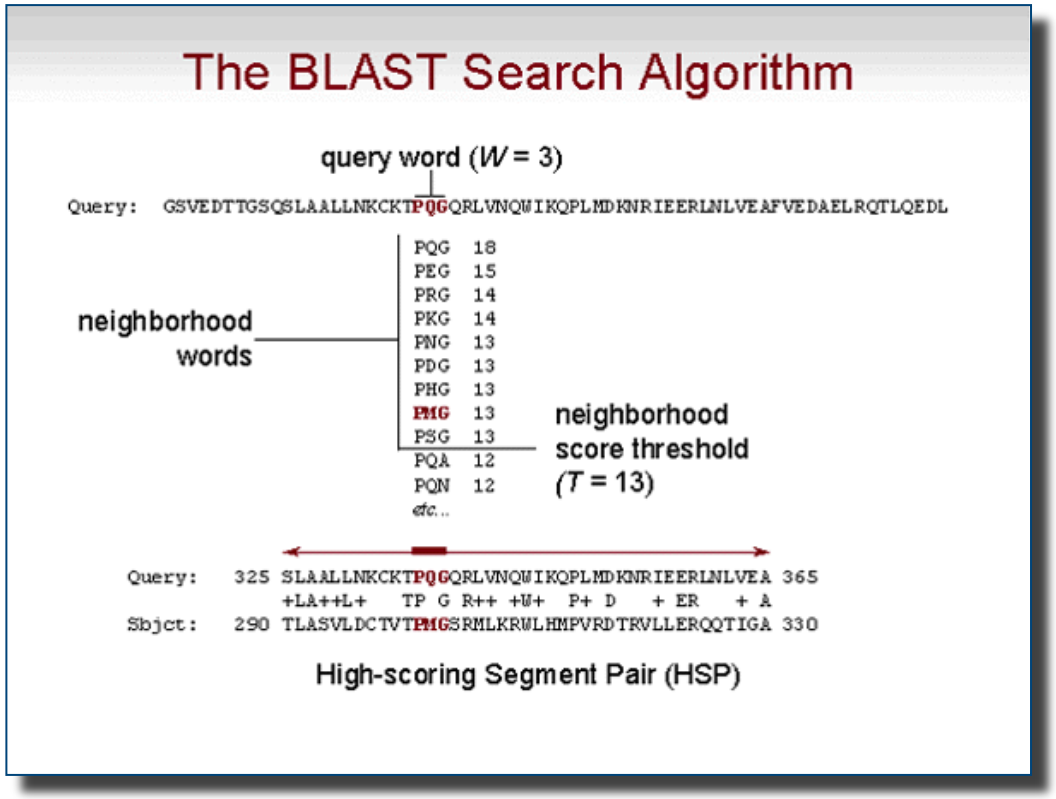
- eine komplette cDNA des Gens bekannt ist
- passende partielle cDNA-Sequenzen (z.B. ESTs) existieren
- auf Proteinebene (also nach Translation der Genom-Sequenz!) eine Ähnlichkeit zu einem bekannten Protein gefunden wird

Vorzugsweise wird zuerst nach Datenbankeinträgen desselben oder naher verwandter Organismen gesucht (auf DNA-Ebene), dann auf Proteinebene nach Ähnlichkeiten in entfernten Organismen (oder entfernt verwandten Proteinen).

Wir erinnern uns....

Der BLAST-Algorithmus

- liefert lokale alignments



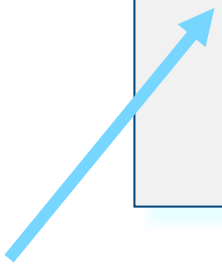
(word size W = 11 bei DNA)

- zunächst wird nach kurzen lokal passenden Abschnitten („words“) gesucht,

dann versucht BLAST2.0, die Bereiche neben den „matching words“ unter Einbeziehung von Lücken zu optimieren

BLAST :
Endecke die Möglichkeiten!

blastn	DNA-Sequenz ÷ DNA-DB > nur nahe Verwandschaft; beide Stränge verglichen
blastp	As-Sequenz ÷ Protein-DB > entfernte Verwandschaft
blastx	DNA-Seq > in 6 Leserahmen translatiert ÷ Protein-DB > findet mögliche Proteine in einer nicht-charakterisierten DNA-Sequenz



BLAST :
Endecke die Möglichkeiten!

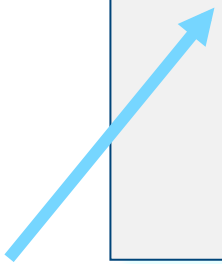
tblastn — As-Seq ÷ DNA-DB (6-frame translatiert!)

> findet nicht-annotierte Genregionen in DNA-DB-Sequenzen

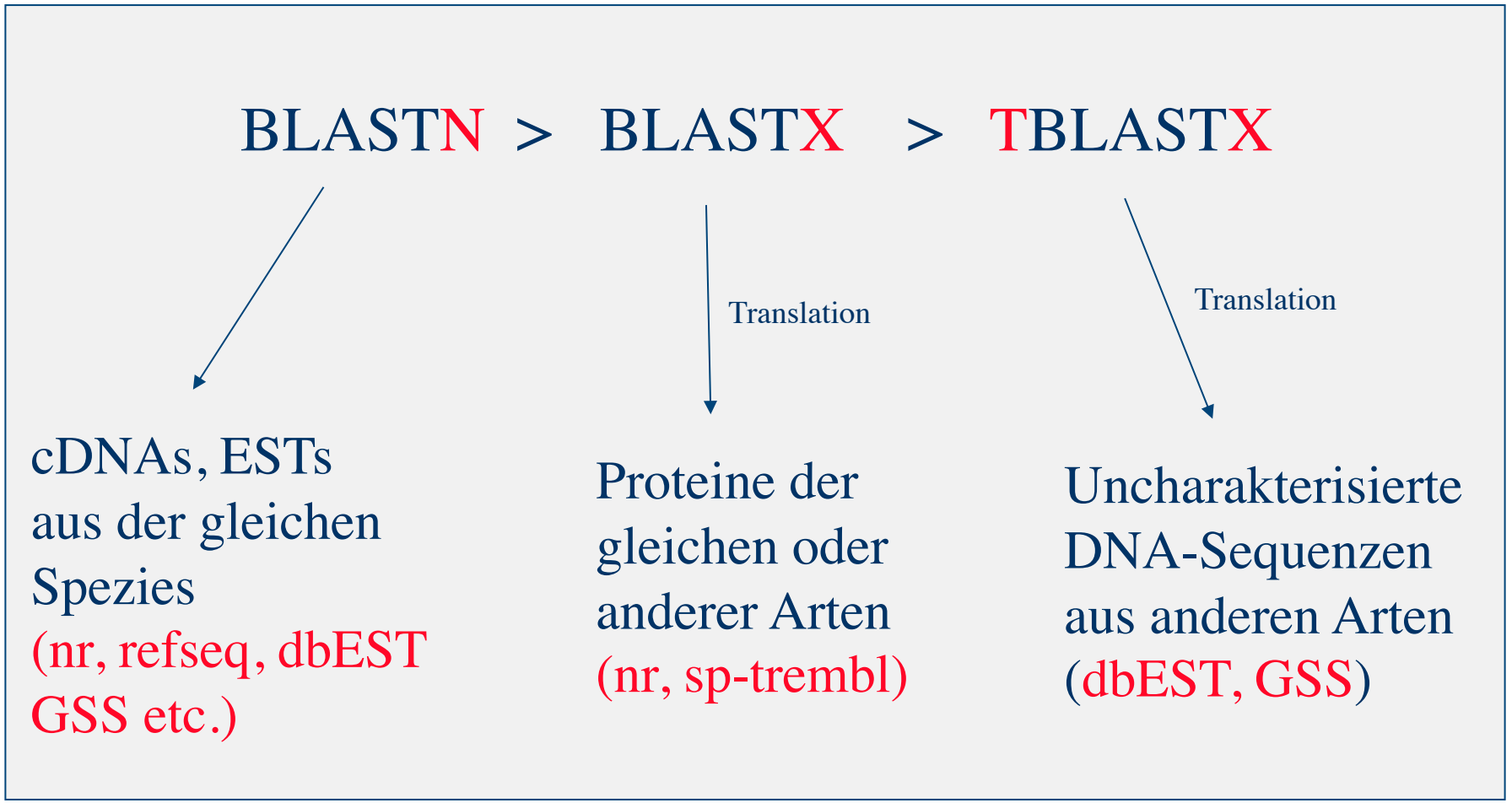
tblastx — 6-frame-Translation einer DNA-Seq ÷ 6-frame-Translation einer DNA-DB

> Analyse z. B. von ESTs auf Proteinebene zur Detektion entfernter Verwandschaften

(kann nicht mit nr-DB benutzt werden (zu aufwändig)



Bei der Suche nach neuen Genen verwenden wir...



Kleiner Exkurs:

Charakterisierung eines gefundenen Gens/Proteins durch BLAST

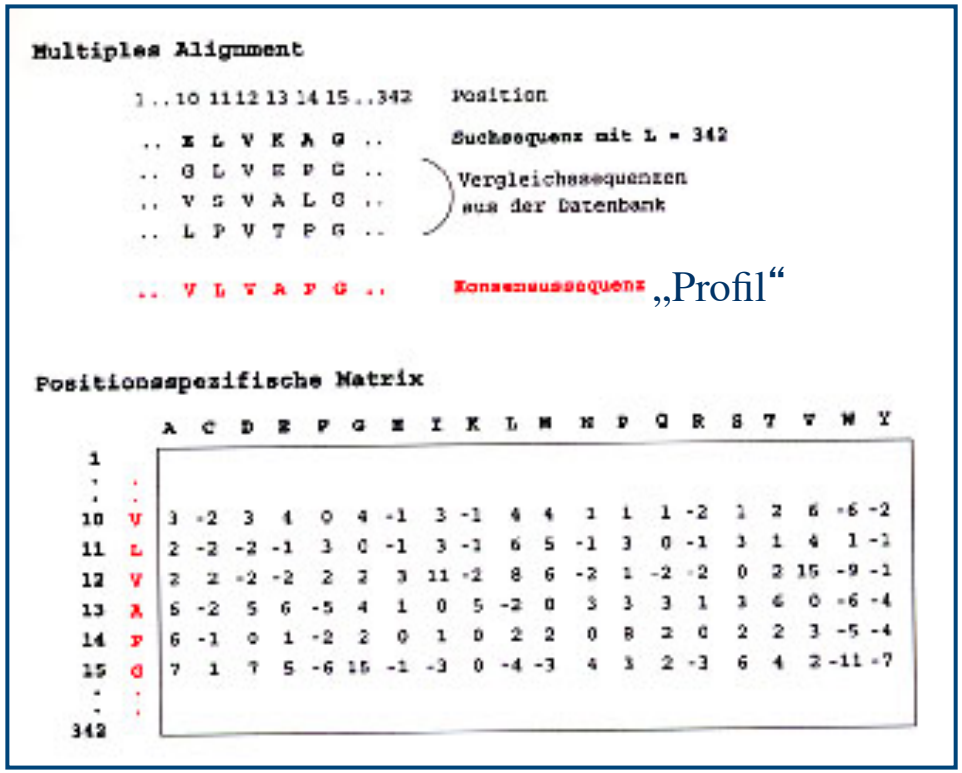
- meist zuerst über BLASTP
- BLASTP findet jedoch keine **sehr** entfernt verwandten Übereinstimmungen mehr.

Hierfür existieren weitere BLAST-Formen...

PSI-BLAST

Position-specific iterated BLAST

- speziell für die Suche sehr entfernt verwandter Proteine



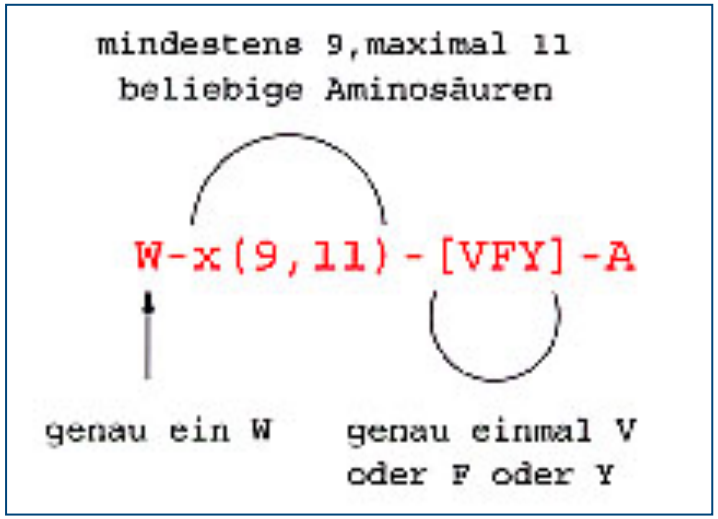
1. **Erste Suche = einfacher BLAST**
2. **Matches untereinander schreiben, > Konsensussequenz errechnen („Profil“)**
3. **„Positions-spezifische“ Substitutions-Matrix errechen**
4. **BLAST mit dem „Profil“ und der PSSM mehrfach wiederholen**

➢ **Gezieltere Suche nach entfernt verwandten Proteinen durch Positionsinformation**

PHI-BLAST

Pattern-hit initiated BLAST

- **sucht Sequenzmuster („Signatur“), das typisch für Proteindomäne ist**
- **Suche über „qualitatives“ Sequenzmotif (PSI-Blast über quantitatives Motiv)**



- **Muster zusammen mit Suchsequenz gegen DB laufen lassen**
- **Treffer = Proteine mit Ähnlichkeit zur Suchsequenz und das Motiv enthaltend**

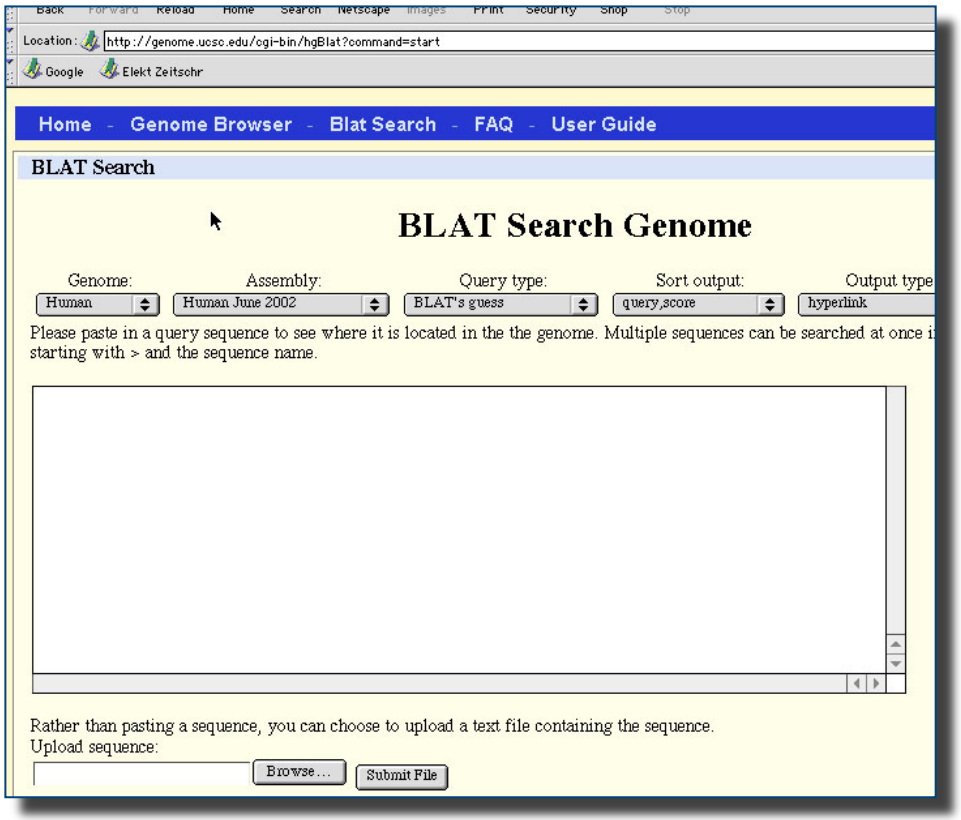
Anwendung: „Suche z.B. alle Proteine mit Zinkfinger-Motiven in der Datenbank“

Zurück zum Problem des „gene findings“ in Genom-DNA per Datenbanksuche...

Am einfachsten ist es natürlich, wenn wir eine passende komplette cDNA (oder zumindest einige ESTs) zu unserem Gen in den Datenbanken finden!

Ultraschnelles Alignment (Genom-DNA/cDNA) über BLAT

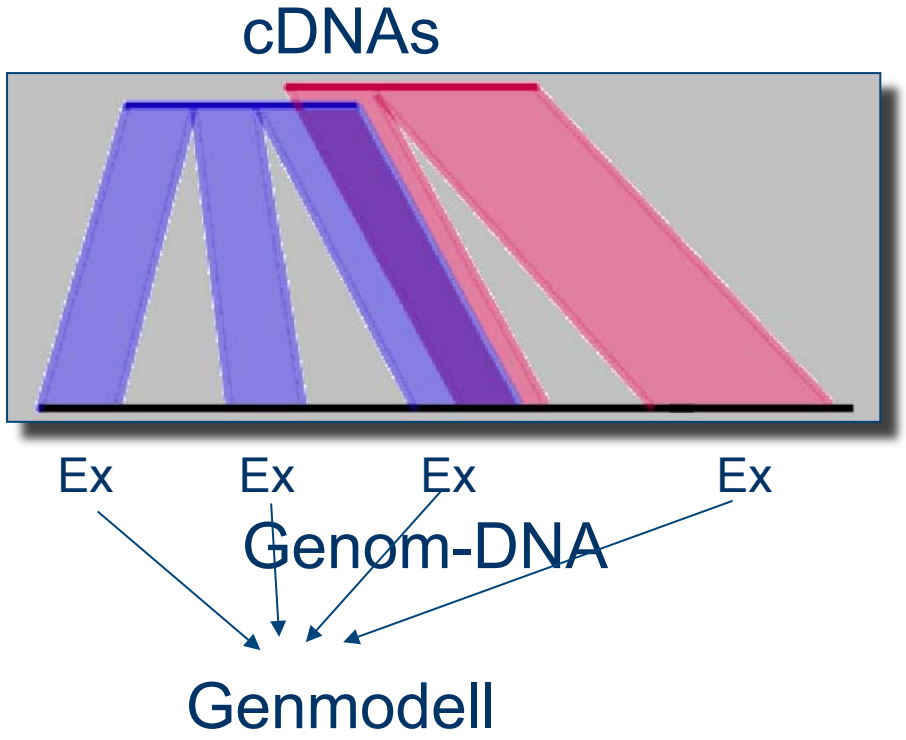
„BLAST-like alignment tool“



http://genome.ucsc.edu/cgi-bin/hgBlat?command=start

- **DNA-BLAT findet 40 Bp (> 95% id) bzw. perfekte matches von >33 Bp**
- **Protein-BLAT findet 20 aa (< 80% id)**
- **Index (DNA) enthält alle nicht-überlappenden 11-mere des Genoms (1 Gb RAM)!!!**
- **Index wird gebraucht um passende Regionen im Genom schnell zu identifizieren, die dann für genaueren Vergleich „hochgeladen“ werden**

„Spliced alignment“ von Genom-sequenz und cDNA



1. **BLAST (mRNA vs. Genom)**
   > Identifizierung des „Genomfensters“
2. **Konsistenz der alignten Abschnitte?**
   (selber DNA-Strang, lineare Abfolge)
3. **„low stringency“-BLAST um komplette mRNA zu alignen**
4. **Anpassen der Lage von Spleiß-Signalen an den Enden der Alignments**

- http://bioweb.pasteur.fr/seqanal/interfaces/est2genome.html

https://www.ncbi.nlm.nih.gov/sutils/splign/splign.cgi

Die drei Wege zum Gen

1. Datenbanksuchen/Alignments

   „es gibt bereits passende Sequenzen in den Datenbanken“

2. „*ab initio*“-Genvorhersage

   „Signale“ in der DNA zeigen: hier ist ein Gen“

3. Vergleich von verwandten Genomen
   („comparative genomics“)

   „hier ist eine evolutionär konservierte Region“

- Genvorhersage macht Modelle für Gene

- Gen-Annotation benutzt diese Modelle, fügt experimentelle Evidenzen hinzu und bewertet das Ergebnis

Edgar Allen Poe und die DNA-Linguistik

Zum Schatz von Captain Kidd... („The Gold-Bug“)

```
53‡‡†305))6*;4826)4‡.)4‡):806*;48†8¶60))8
5;1‡(;:‡*8†83(88)5*†;46(88*96*?;8)*‡(;485
);5*†2:*‡(;4956*2(5*--4)8¶8*;4069285);)6†8
)4‡‡;1(‡9;4808 1;8:8‡1;48†85;4)485†528806
*81(‡9;48;(88;4(‡?34;48)4‡;161;:188;‡?;
```

- häufigstes engl. Wort? ;48 the

```
53‡‡†305))6*THE26)H‡.)H‡):E06*THE†E¶60))E
5T1‡(T:‡*E†E3(EE)5*†T46(EE*96*?TE)*‡(THE5
)T5*†2:*‡(TH956*2(5*--H)E¶E*TH0692E5)T)6†E
)H‡‡T1(‡9THE0E1TE:E‡1THE†E5TH)HE5†52EE06
*E1(‡9THET(EETH(‡?3HTHE)H‡T161T:1EET‡?T
```

Pevzner, P. 2000

Welche „Signale“ von Genen kennen wir?

- Repeats = meist keine Gene > also filtern/maskieren

- Startkodons, Stopkodons > ORFS („open reading frames“)
- Spleiß-Donor/Akzeptor-Stellen (“GT-intron-AG“)

- Promoter: Bindemotive für Transkriptionsfaktoren („Boxen“)
  Startpunkt der Transkription (+1, cap site)
  CpG-Inseln

- Polyadenylierungssignal (AATAAA) am Ende des Transkripts

Voraussetzung für viele weitere Analysen:

Maskieren von Repetitionen

```
 SW   perc perc perc   query         position in query    matching repeat        position in  repeat
score  div. del. ins.   sequence      begin  end (left)   repeat   class/family begin  end (left)   ID

 1580   8.7  2.3 10.0  UnnamedSeq1      38  507    (12) +  SVA       Other            433  866  (520)    1


Masked Sequence:


>UnnamedSeq1
CGCCCGGCCAGCCGCCCCGTCCGGGAGGTGAGGGGCTNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNGTTGGGGGGGGG
```

„REPEATMASKER“

http://ftp.genome.washington.edu/cgi-bin/RepeatMasker

Probleme:

- Repeats sind artspezifisch
- Repeats mutieren rasch zur Unkenntlichkeit und werden nicht maskiert

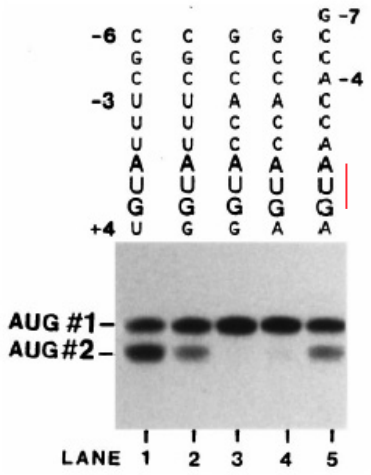
Manche „Signale“ sind recht eindeutig definiert...

Bsp. Startkodon

Kozak‘s rule“:

Meist gilt das erste „in frame“-ATG vom 5`-Ende der mRNA her gesehen als Startkodon

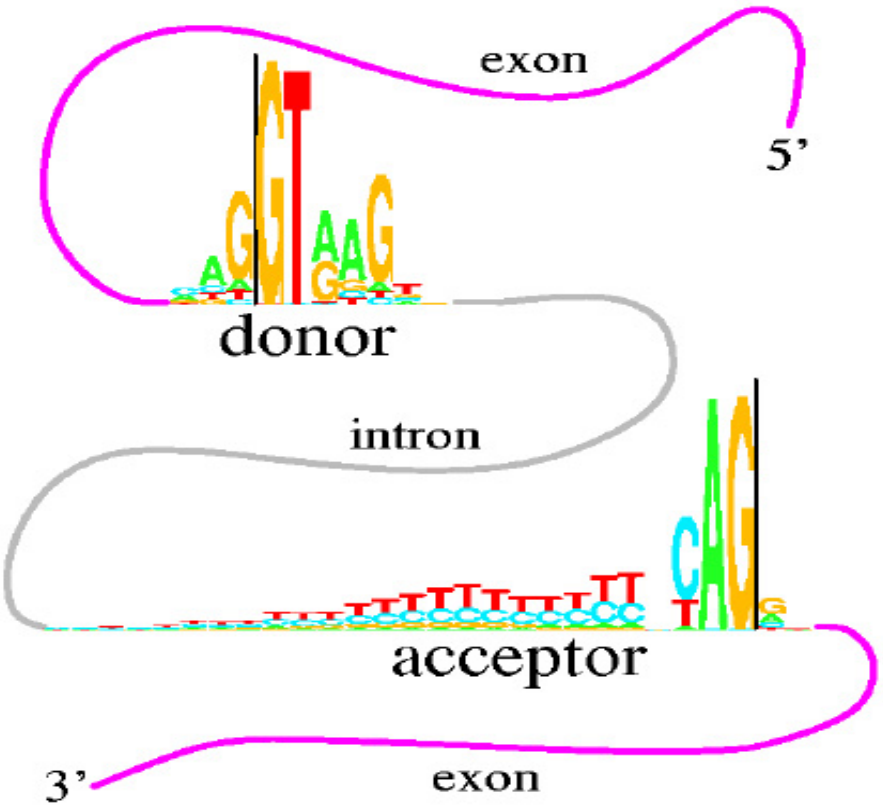
Aber es gibt Ausnahmen:



Gene 234 (1999) 187-208

Viele „Signale“ sind jedoch unscharf definiert...

Bsp. Spleißstellen



• was unterscheidet eine effiziente von einer kryptischen Spleiß-Stelle??

„Position-specific weight matrices“

erkennen Signale besser als Konsensus-Sequenzen

Problem von Konsensus-Sequenzen:

```
MELON
MANGO
HONEY
SWEET
COOKY
-----
MONEY
```

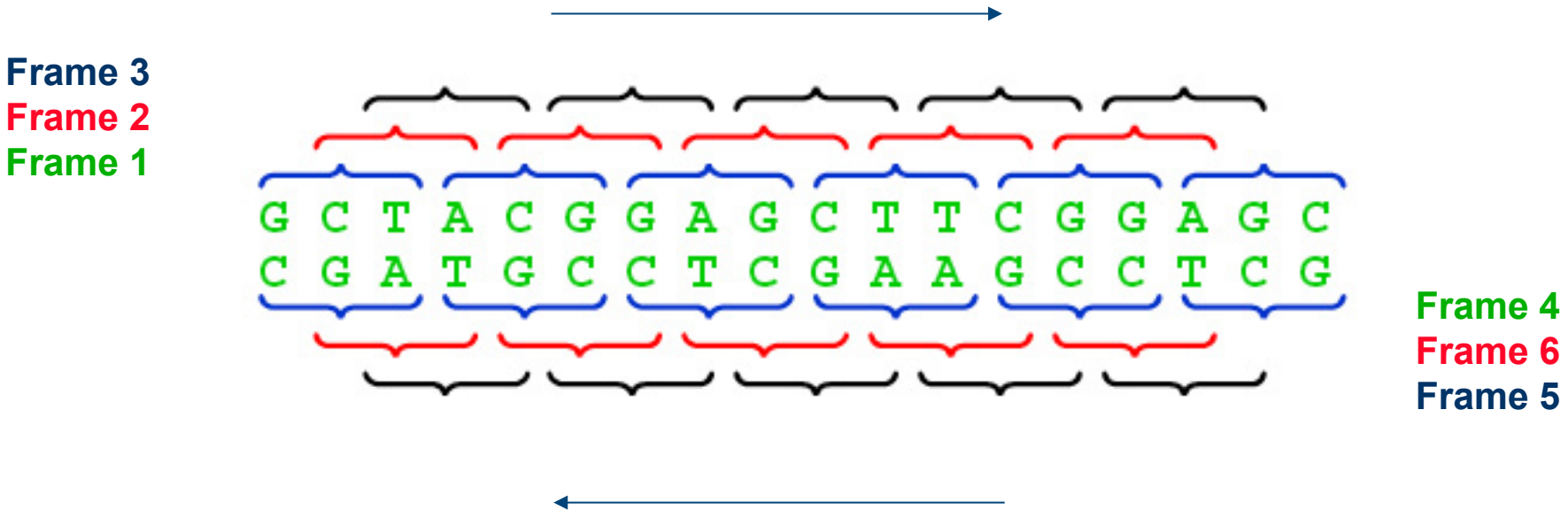
Bsp. „branch site“ von Introns

- PSWM:

	-3	-2	-1	0	+1
A	1	0	39	**99**	4
C	76	8	15	1	45
G	2	0	42	0	6
T	21	91	4	0	38

Proteinkodierende Gene haben einen „besonderen Inhalt“

- sie lassen sich als einen „offenen Leserahmen“ (ORF) lesen

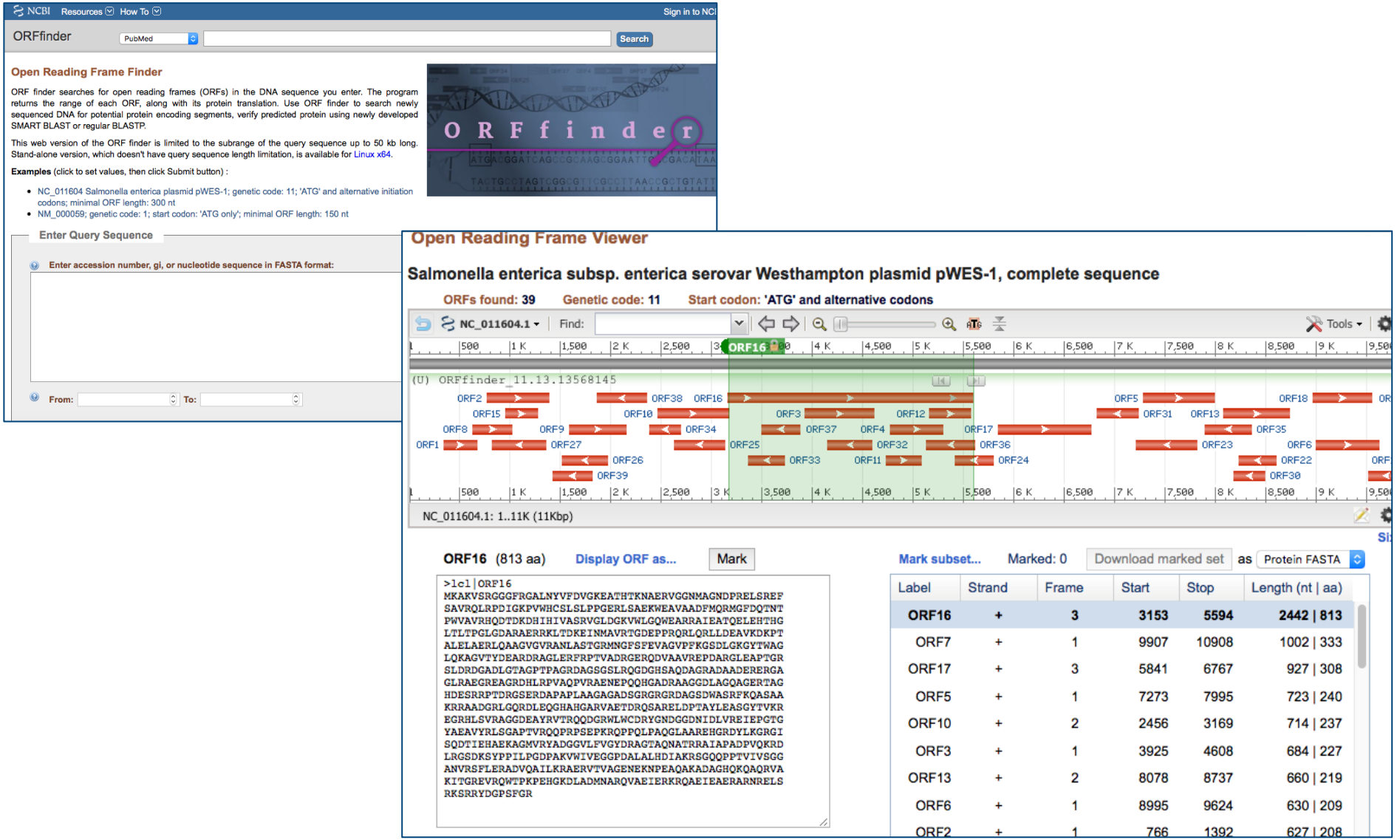


Suche nach ORFs

Start
Stop

Offene Leserahmen (ORFs)

Der NCBI-ORFfinder



Proteinkodierende Gene haben einen „besonderen Inhalt“

- ihre Nukleotid-Komposition unterscheidet sich von nicht-kodierender DNA:

  > bestimmte Aminosäuren/Codons sind häufiger als andere (z. B. ist Leu häufiger als Trp)

  > es gibt unterschiedliche Kodonanzahlen für verschiedene As (Leu:6 Trp:1)

  > für eine bestimmmte As werden bestimmte Kodons häufiger gebraucht als andere („codon usage“)

Proteinkodierende Gene haben einen „besonderen Inhalt“

- Di-Nukleotid-Häufigkeiten: z.B. „AT“ in proteinkodierenden Sequenzen unterrepräsentiert

Sequenzbeispiel:
M. Stanke, Greifswald

Codon usage

- Kodons werden unterschiedlich häufig verwendet
- dies kann sogar je nach Spezies unterschiedlich sein
  (> artspezifische Trainingsdatensätze erforderlich!)

Bsp. E. coli

AA	codon	/1000
Gly	GGG	1.89
Gly	GGA	0.44
Gly	GGU	52.99
Gly	GGC	34.55
Glu	GAG	15.68
Glu	GAA	57.20
Asp	GAU	21.63
Asp	GAC	43.26

Eine codon-usage-Methode zur Erkennung einer proteinkodierenden Sequenz

- Staden and Mc Lachlan 1982:

Bestimme die Wahrscheinlichkeit P, dass eine Sequenz S „so aussieht“, **weil** sie ein Protein kodiert

- Die Sequenz in einem Sequenzfenster fällt immer in eine von 7 Kategorien (frame 1...frame 6, non-cod)
- Benutze Bayes-Satz, um die Wahrscheinlichkeit jeder Kategorie zu bestimmen:

$$\Pr(\text{coding}_i \mid S) = \frac{\Pr(S \mid \text{coding}_i)\Pr(\text{coding}_i)}{\Pr(S)}$$

Wahrscheinlichkeit, dass Sequenz ein Protein in frame i kodiert

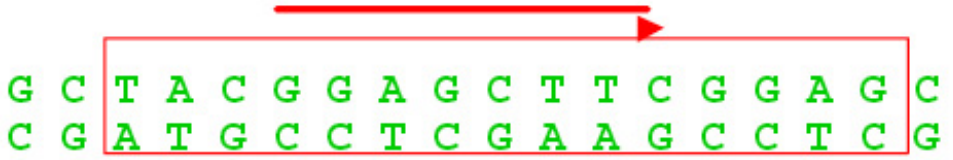
Eine codon-usage-Methode zur Erkennung einer proteinkodierenden Sequenz

- Kalkuliere: $$\Pr(S \mid \text{coding}_i) \approx \prod_{j=1}^{n} \Pr(S_i(j) \mid \text{coding}_i)$$

Wahrscheinlichkeit des Kodons j in frame i, gegeben die Sequenz ist kodierend

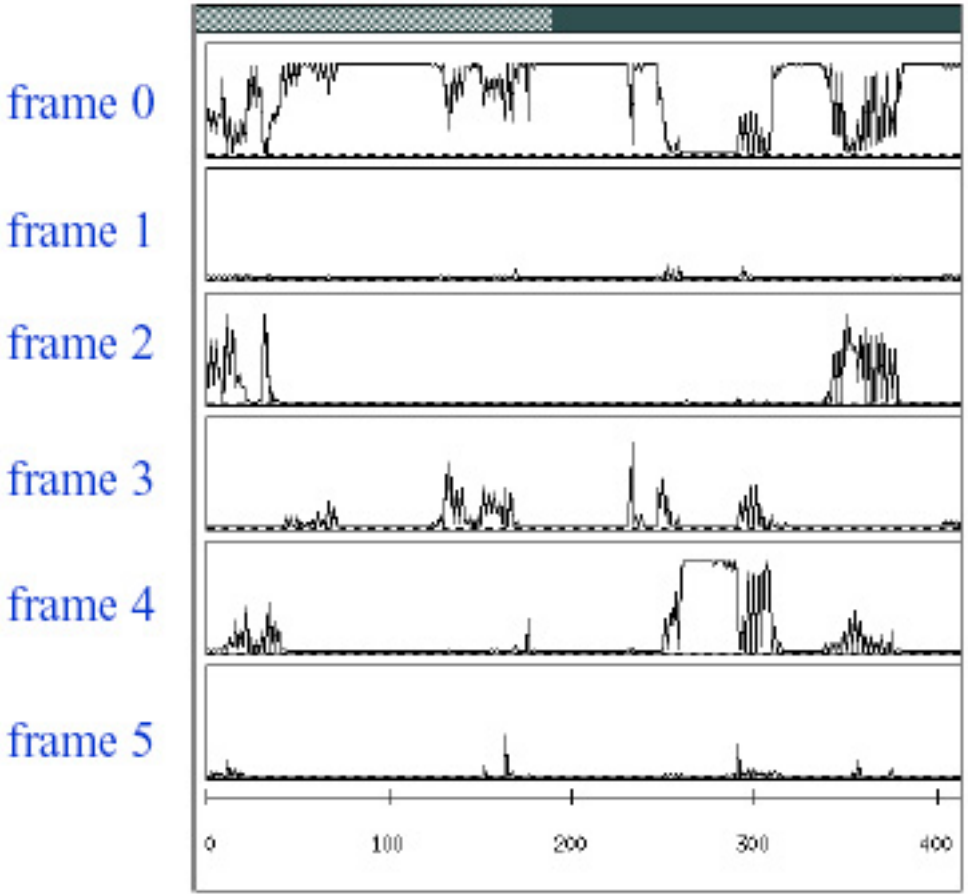
G C T A C G G A G C T T C G G A G C
C G A T G C C T C G A A G C C T C G

$$\Pr(S \mid \text{coding}_0) \approx$$
$$\Pr(\text{GCT} \mid \text{coding}_0) \times \Pr(\text{ACG} \mid \text{coding}_0) \ ..$$



Sliding window

Eine codon-usage-Methode zur Erkennung einer proteinkodierenden Sequenz



- aber: recht grobe Analyse der Kodierungskapazität

➢Exakte Begrenzung der ORFs nicht festgelegt

Bisherige Suchen sind nicht ausreichend, um komplizierte Genmodelle exakt vorherzusagen!

Moderne ***integrierte*** Genvorhersage-Programme verbinden Suche nach Signalen und Inhalt mit statistischem Vorab-Wissen über Gene...

...Hidden Markov Models (HMM)

Markov WER??

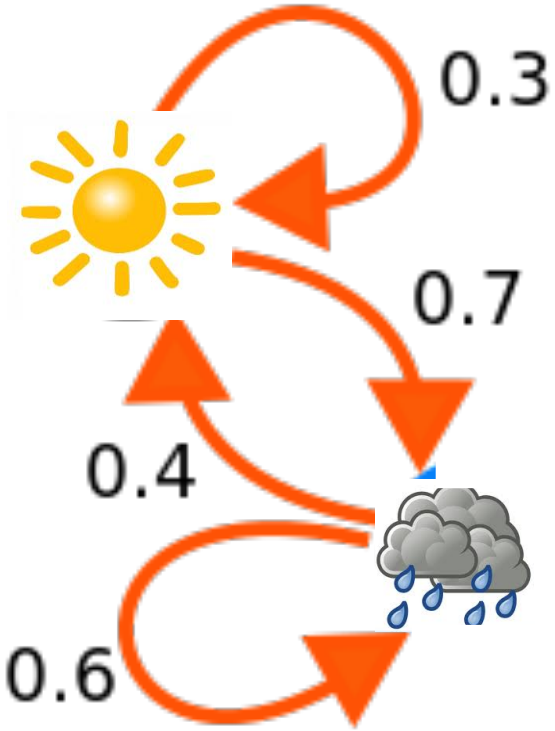
- Andrei Andreyevich Markov (1856-1922)
- Markov-Kette:

Eine *Markovkette* ist ein stochastischer Prozess, der nacheinander eine Reihe von Zuständen mit einer gewissen Wahrscheinlichkeit durchläuft. Dabei hängt die Wahrscheinlichkeit für den jeweils nächsten Zustand nur vom aktuellen Zustand ab:

$$P(t_{i+1} | t_i, t_{i-1}, \ldots, t_j) = P(t_{i+1} | t_i)$$

Pfeile geben Übergangswahrscheinlichkeiten an

Markov-Kette „Wettervorhersage“



Hidden Markov Models

- verwende **statistische Informationen,** um Abfolgen (z. B. Sequenzen) zu klassifizieren

- Analogie:

  „Automatische Erkennung der Sprache eines Textes“

  In einem typischen deutschen Text macht der Buchstabe ‚e‘ ca. 16,55% aller Buchstaben aus, in einem schwedischen nur ca. 9.77%.

  ⇒ zähle die e‘s im Text, um zu berechnen mit welcher Wahrscheinlichkeit es sich um einen deutschen Text handelt

Hidden Markov Models

Was ist denn da „hidden“??

- wir ***sehen*** nur die **„e ‘s“** — **„emission“**
- dahinter ***versteckt*** sich die Information:

„dies ist ein deutscher Text“ — **„state“**

Illustration aus der Informatik-Welt...

„The Occasionally Dishonest Casino“

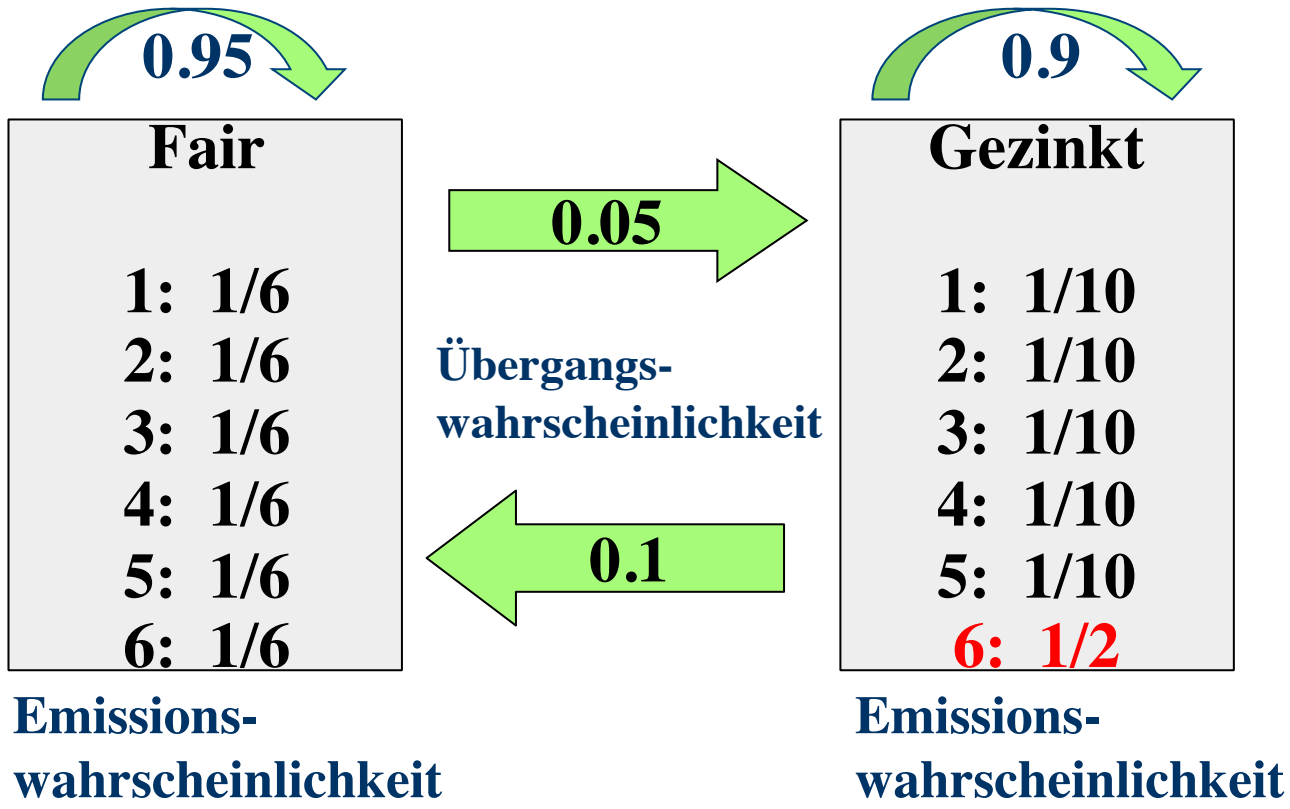
- Das Casino besitzt gezinkte und faire Würfel
- Der gezinkte Würfel hat 50% Chance eine 6 zu bekommen und gleiche Chancen (10%) eine der anderen Zahlen zu bekommen.
- Das Casino wechselt von Zeit zu Zeit die Würfel, um die Spieler zu betrügen.

```
Zahlen:   315116246446644245311321363165662656666651166453132651245
```

Wann ist der Würfel gewechselt worden?

The Occasionally Dishonest Casino

Annahme: wir kennen die Wahrscheinlichkeit des Würfel-Wechsels und die Wahrscheinlichkeiten der von diesen Würfeln gewürfelten Zahlen



The Occasionally Dishonest Casino

- Der Viterbi-Algorithmus (Viterbi 1967) berechnet die Würfelsituation auf Basis der gewürfelten Zahlen.
- Das HMM ist die Grundlage des Algorithmus.

```
Zahlen:    3151162464466442453113213631656626566666511664531326512 45
Würfel:    FFFFFFFFFFFFFFFFFFFFFFFFGGGGGGGGGGGGGGGGGGGFFFFFFFFFFFF
Viterbi:   FFFFFFFFFFFFFFFFFFFFFFFFFFFFGGGGGGGGGGGGGGGGGGFFFFFFFFFFFF
```

Hidden Markov Models

- Anwendungsgebiete in der Bioinformatik:

  > Vorhersage der Genstruktur (Exons/Introns)

  > Vorhersage von Promoterbereichen

  > Erstellung von Modellen für Proteinfamilien zum Suchen nach entfernt verwandten Proteinen in DB („profile HMMs“)

Von der Textsuche zum HMM

1	ACA---ATG
2	TCAACTATC
3	ACAC--AGC
4	AGA---ATC
5	ACCG--ATC

Bsp.: Fünf Sequenzen, die ein Signal definieren

Einfache Textsuche würde erfolgen nach:

`(AT)(GC)(AC)(ACGT)* A(TG)(GC)`

Dies kann bei Suche nicht unterscheiden zwischen...

...einer plausiblen Sequenz (z.B. der Konsensus-S.)

`ACAC--ATC`

...und einer höchst unwahrscheinlichen Sequenz

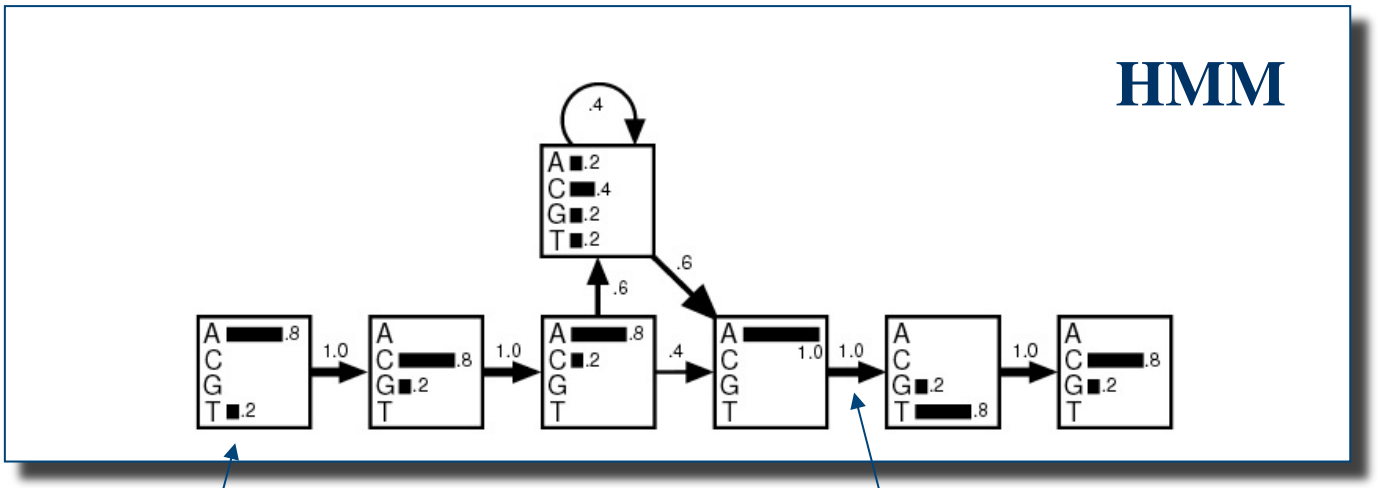
`TGCT--AGG`

HMM

Also besser:

Bewerten, ob Sequenzabfolge „gut“ in das Alignment passt...

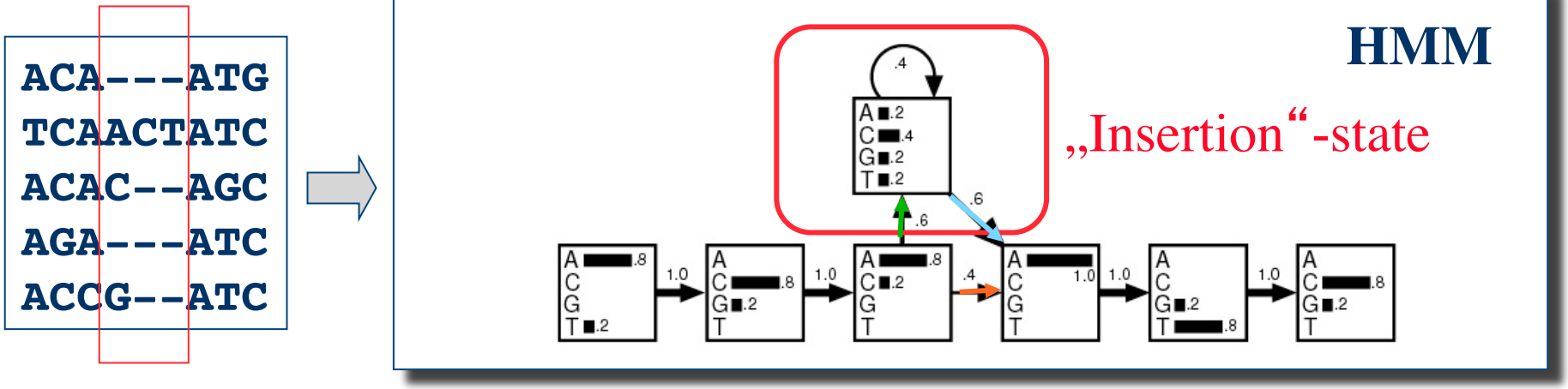
```
ACA---ATG
TCAACTATC
ACAC--AGC
AGA---ATC
ACCG--ATC
```



Emissions-
wahrscheinlichkeit

Übergangs-
wahrscheinlichkeit

HMM

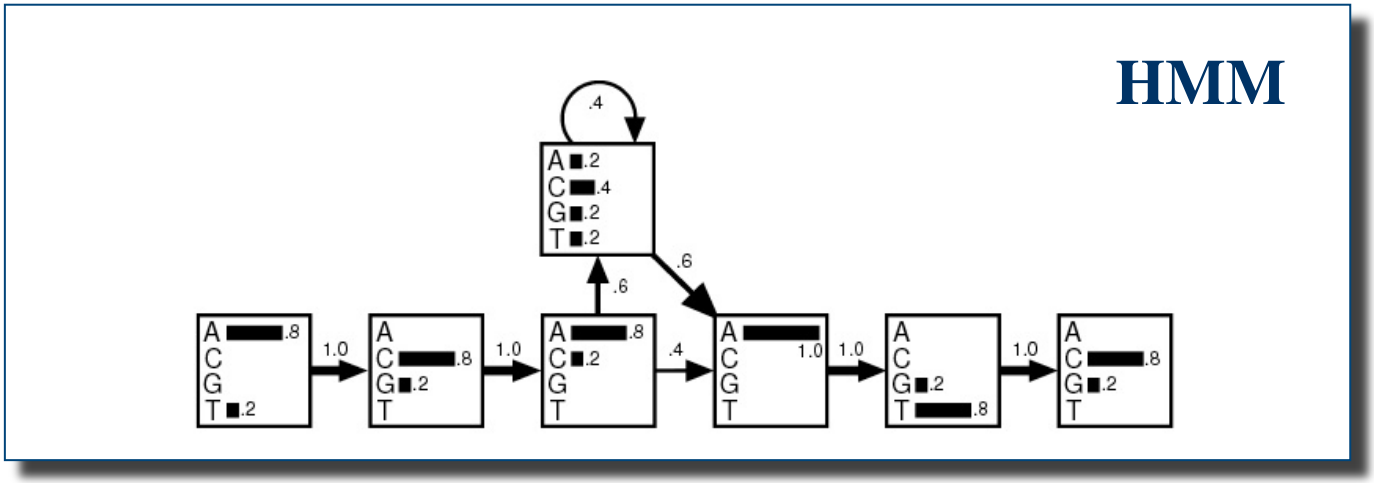


- nach Pos. 3 haben 3/5 Sequenzen eine Insertion (=0.6)
- diese 3/5 (=0.6) Sequenzen gehen von Insertionstate zu Pos. 4
- 2/5 (=0,4) Sequenzen gehen von Pos. 3 zu Pos.4

- nach der ersten Insertion (Sequ 2,3,5) kehrt Sequ 2 zweimal (=0.4) in den Insertion-State zurück

HMM

```
ACA---ATG
TCAACTATC
ACAC--AGC
AGA---ATC
ACCG--ATC
```



Jetzt bewerten wir damit als Beispiel die Konsensussequenz:

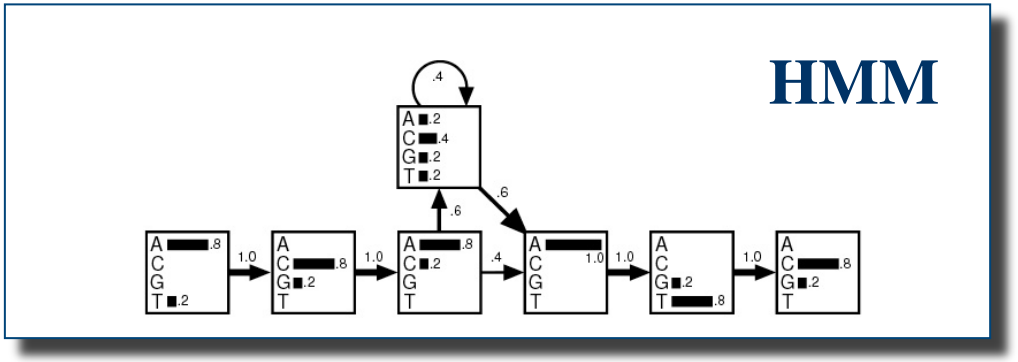
ACAC--ATC

$$
\begin{aligned}
P(\texttt{ACACATC}) &= 0.8 \times 1 \times 0.8 \times 1 \times 0.8 \times 0.6 \times \\
&\quad 0.4 \times 0.6 \times 1 \times 1 \times 0.8 \times 1 \times 0.8 \\
&\simeq 4.7 \times 10^{-2}.
\end{aligned}
$$

Die „unplausible“ Sequenz `TGCT--AGG` ergibt: 0,0023 X 10^{-2}

HMM

```
ACA---ATG
TCAACTATC
ACAC--AGC
AGA---ATC
ACCG--ATC
```



Bewertung aller Sequenzen des Alignments:

	Sequence	Probability ×100	Log odds
Consensus	A C A C - - A T C	4.7	6.7
Original sequences	A C A - - - A T G	3.3	4.9
	T C A A C T A T C	0.0075	3.0
	A C A C - - A G C	1.2	5.3
	A G A - - - A T C	3.3	4.9
	A C C G - - A T C	0.59	4.6
Exceptional	T G C T - - A G G	0.0023	-0.97

HMM

	Sequence	Probability ×100	Log odds
Consensus	A C A C - - A T C	4.7	6.7
Original	A C A - - - A T G	3.3	4.9
sequences	T C A A C T A T C	0.0075	3.0
	A C A C - - A G C	1.2	5.3
	A G A - - - A T C	3.3	4.9
	A C C G - - A T C	0.59	4.6
Exceptional	T G C T - - A G G	0.0023	-0.97

Da der P-Wert stark von der Länge der Sequenz abhängt, erfolgt eine Normalisierung der P-Werte auf die Sequenzlänge (und bessere Skalierung durch Umrechnung in „log odds“-Werte)

$$\text{log-odds for sequence } S = \log \frac{P(S)}{0.25^L} = \log P(S) - L \log 0.25.$$

Wahrscheinlichkeit einer DNA-Sequenz der Länge L

Diskriminierung passender Sequenzen gegenüber schlechten ist nach Normalisierung und Umrechnung sehr viel besser: vergleiche Sequenz 2 im Alignment (s.o.)

HMM

Die Umrechnung von P-Werten auf Log odds-Werte erfolgt primär aus mathematischen Gründen :

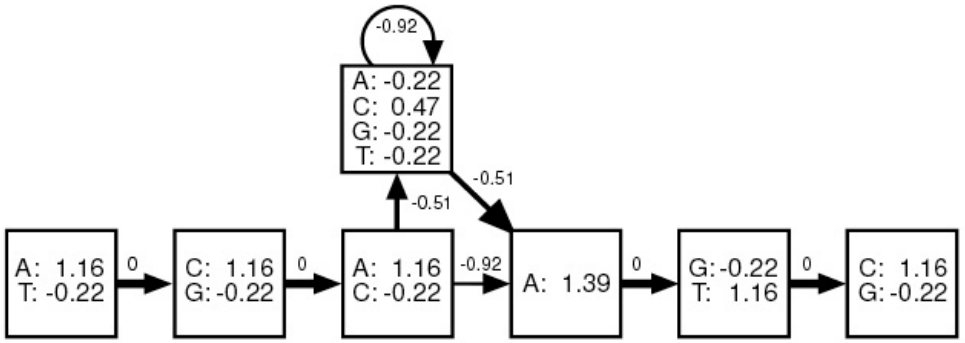


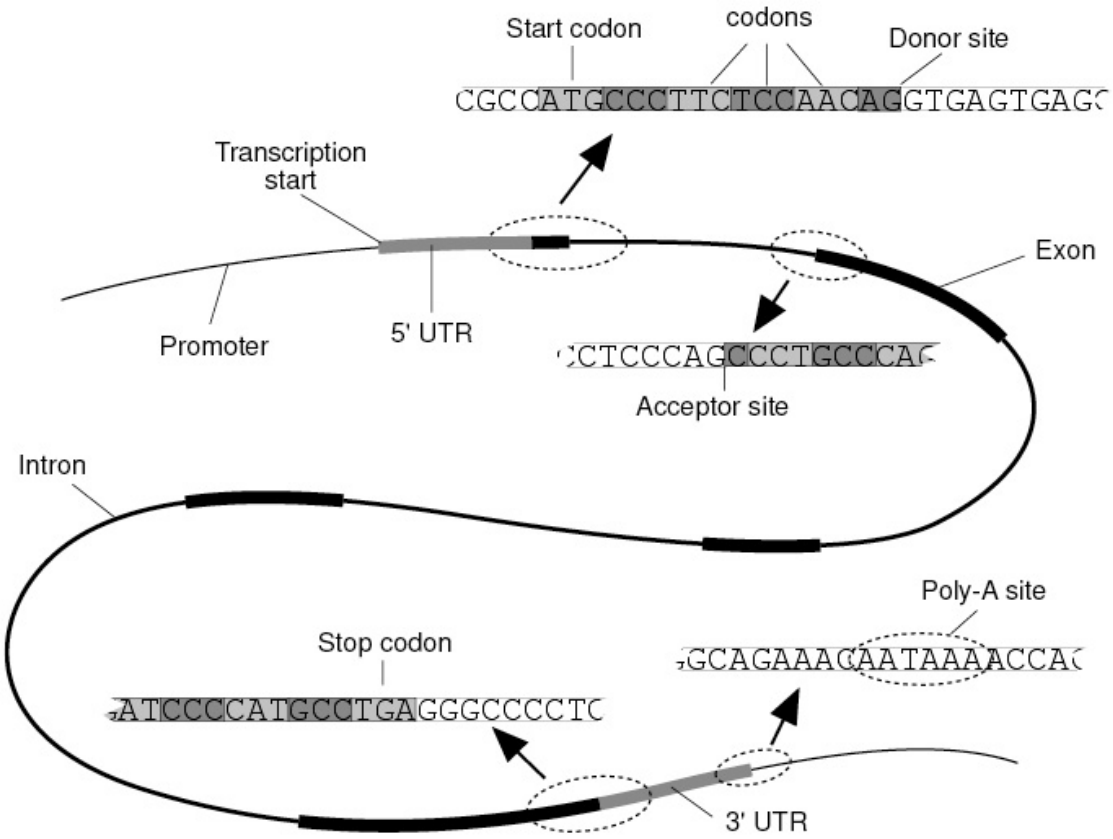
Figure 4.2: The probabilities of the model in Fig. 4.1 have been turned into log-odds by taking the logarithm of each nucleotide probability and subtracting log(0.25). The transition probabilities have been converted to simple logs.

Nun ist eine einfache Summation möglich:

$$\begin{aligned} \text{log-odds}(\texttt{ACACATC}) &= 1.16+0+1.16+0+1.16-0.51+ \\ &\quad 0.47-0.51+1.39+01.16+0+1.16 \\ &= 6.64. \end{aligned}$$

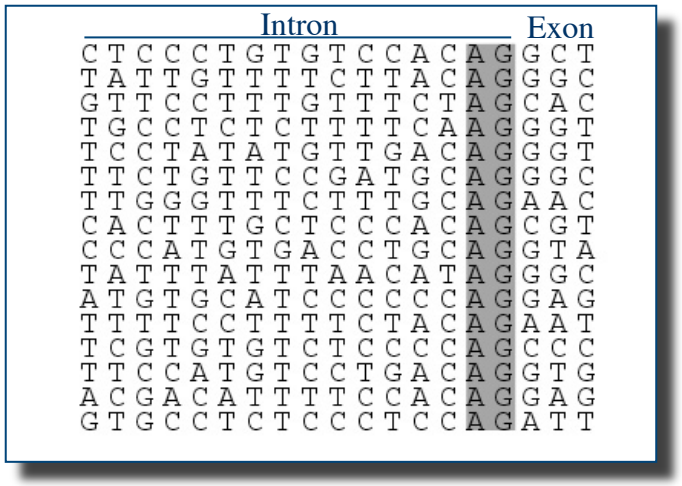
HMMs zur Genvorhersage

Finde die „Grammatik“ des Genoms...



HMMs zur Genvorhersage

Beispiel: ein HMM-Sensor für die 3 ‘Spleißstelle



Bisheriges HMM war zu einfach, da in der DNA starke Präferenzen für **Di**-Nukleotide bekannt sind!

d.h., ein Nt in der Folge hängt stark von seinem Nachbarn ab

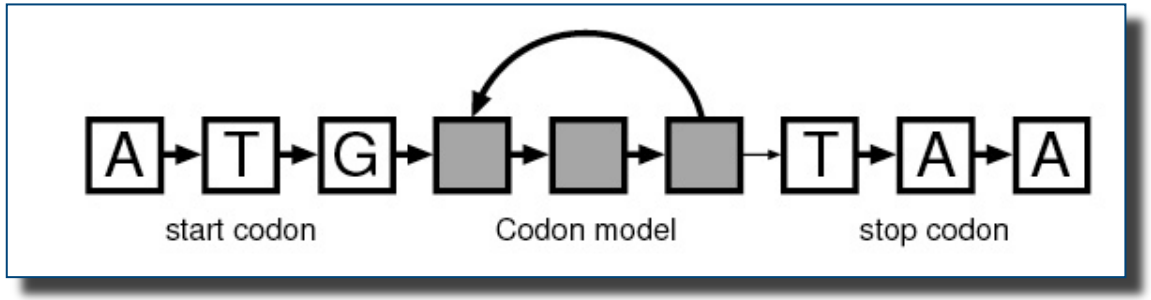
Wir müssen konditionale Wahrscheinlichkeiten für das HMM kalkulieren:

$$P(ACTGTC\ldots) = p_1(A) \times p_2(C|A) \times p_3(T|C) \times p_4(G|T) \times p_5(T|G) \times p_6(C|T) \times \ldots$$

Benutzung konditionaler Wahrscheinlichkeiten > HMM „erster Ordnung“

HMMs zur Genvorhersage

Beispiel: ein HMM-Modell für ein ungespleißtes proteinkodierendes Gen



HMM zweiter Ordnung:
Kodonpos. 3 ist abhängig von Pos. 1 und 2

- zähle die Anzahl jedes der 64 Kodons in einem Datensatz von bekannt-kodierenden Genregionen aus und berechne deren Wahrscheinlichkeiten

z.B. für Kodons CAA,CAC,CAG und CAT

$$
\begin{aligned}
p(\mathrm{A}|\mathrm{CA}) &= c(\mathrm{CAA})/[c(\mathrm{CAA})+c(\mathrm{CAC})+c(\mathrm{CAG})+c(\mathrm{CAT})] \\
p(\mathrm{C}|\mathrm{CA}) &= c(\mathrm{CAC})/[c(\mathrm{CAA})+c(\mathrm{CAC})+c(\mathrm{CAG})+c(\mathrm{CAT})] \\
p(\mathrm{G}|\mathrm{CA}) &= c(\mathrm{CAG})/[c(\mathrm{CAA})+c(\mathrm{CAC})+c(\mathrm{CAG})+c(\mathrm{CAT})] \\
p(\mathrm{T}|\mathrm{CA}) &= c(\mathrm{CAT})/[c(\mathrm{CAA})+c(\mathrm{CAC})+c(\mathrm{CAG})+c(\mathrm{CAT})]
\end{aligned}
$$

C = Codonanzahl

HMMs zur Genvorhersage

Kombination der einzelnen Sensoren:

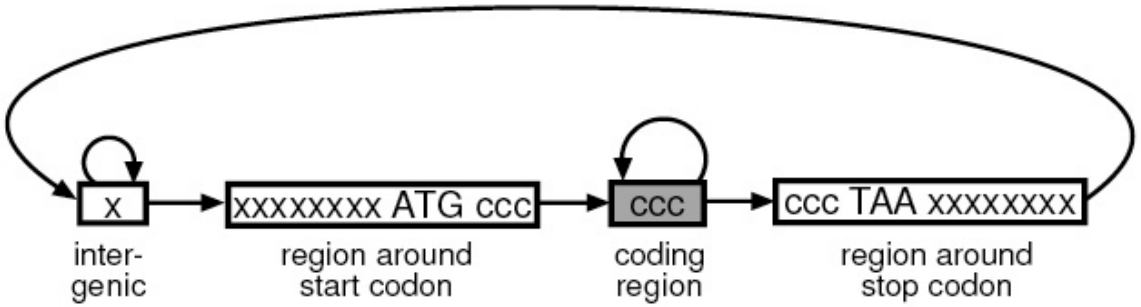
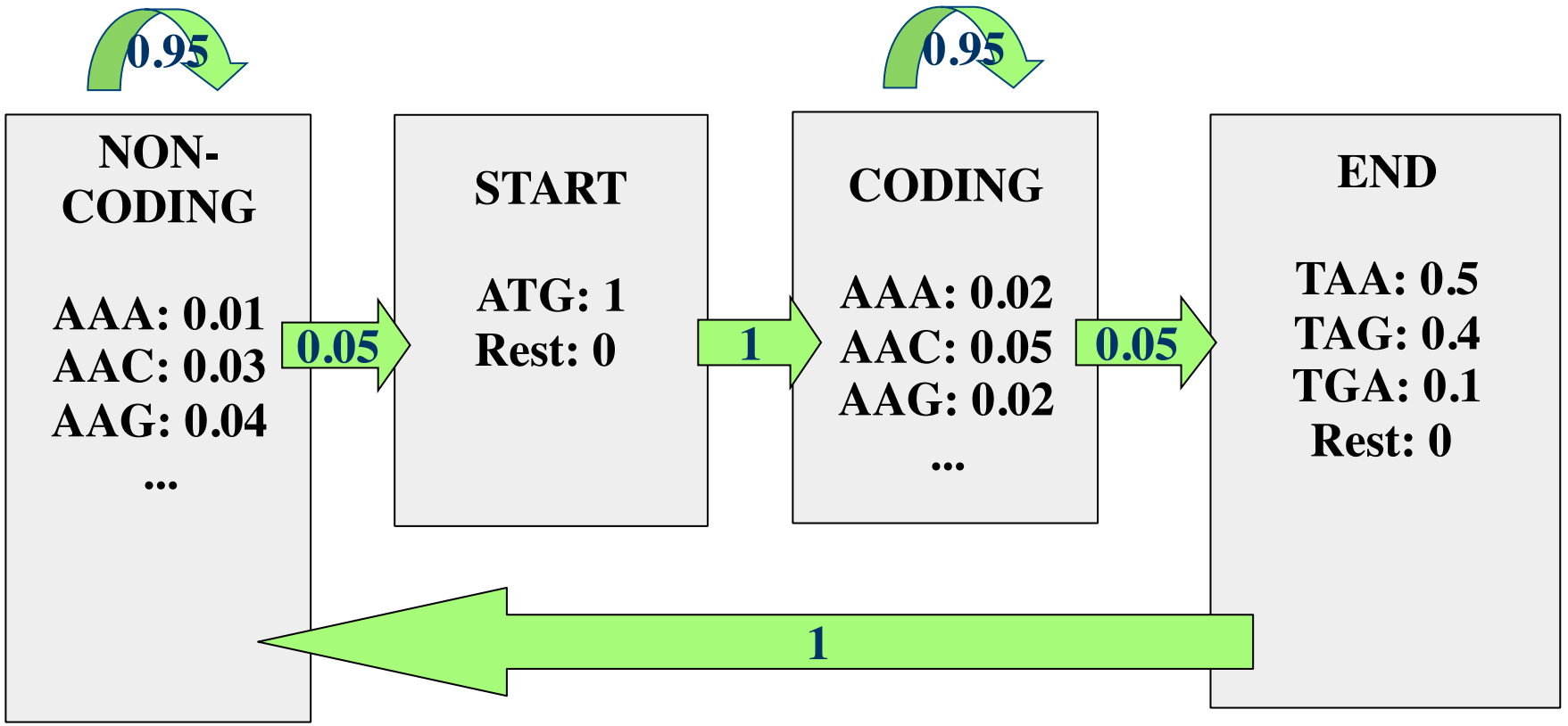


Figure 4.11: A hidden Markov model for unspliced genes. In this drawing an 'x' means a state for non-coding DNA, and a 'c' a state for coding DNA. Only one of the three possible stop codons are shown in the model of the region around the stop codon.

HMM für prokaryotisches Gen

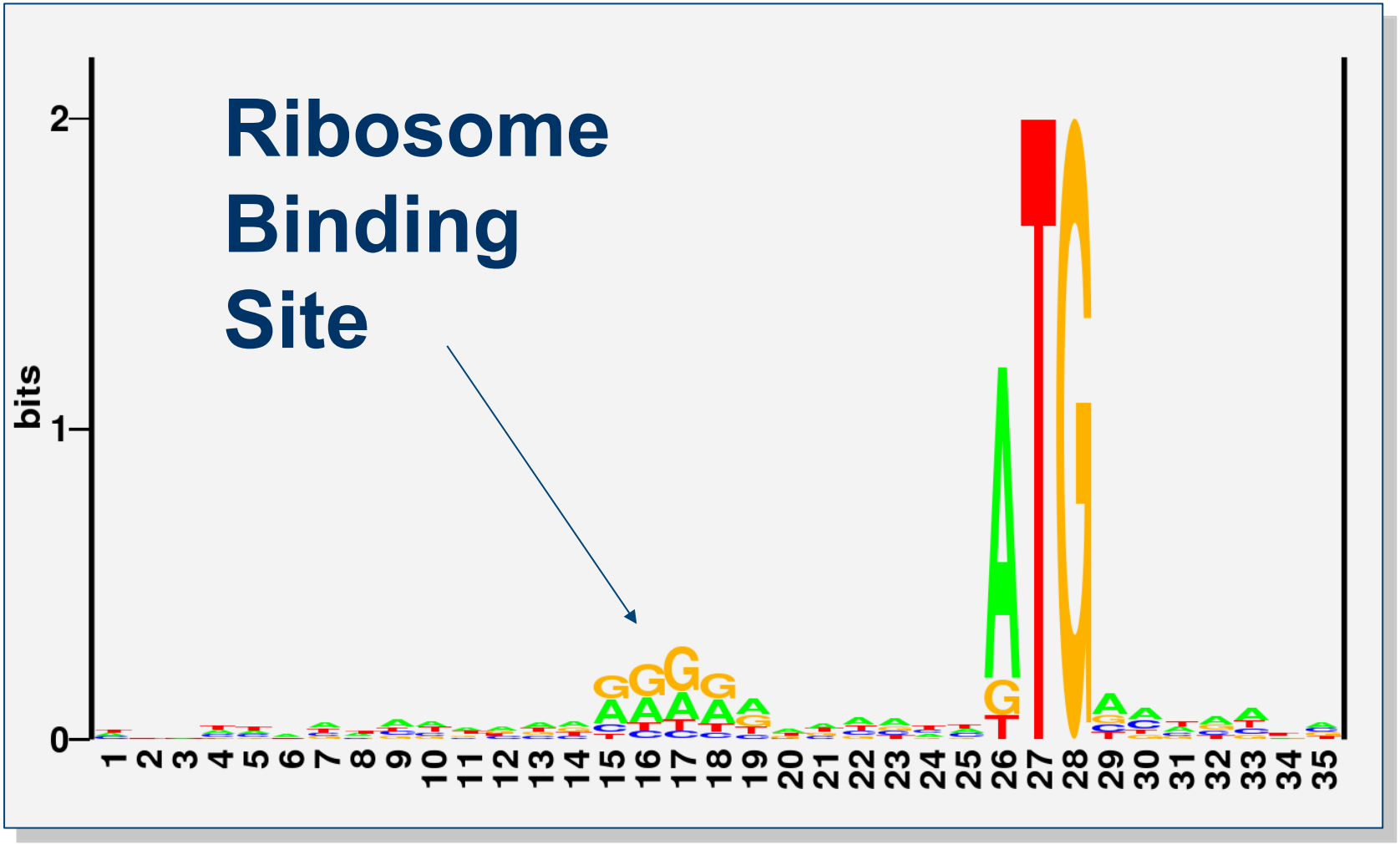


HMM für prokaryotisches Gen

Ein noch relativ übersichtliches Problem….

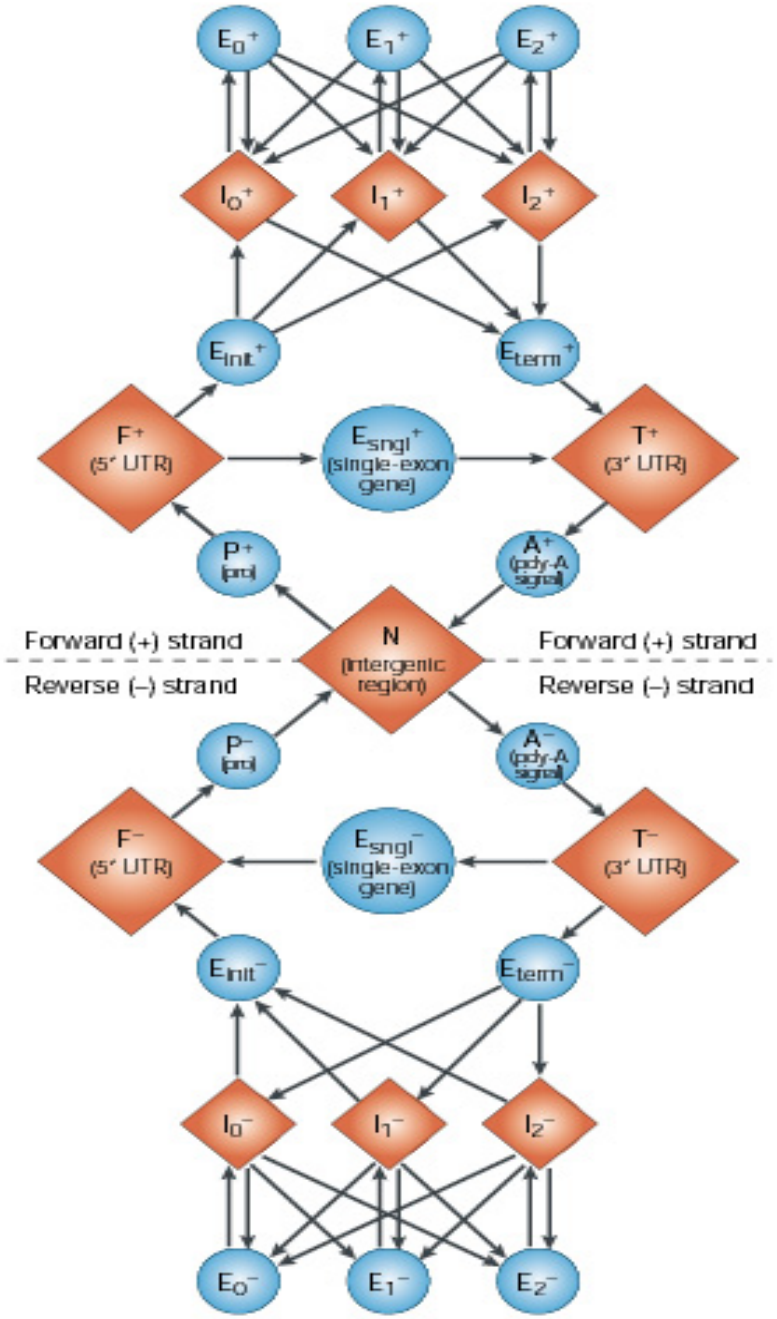
- **Kleine Genome (0.5-10 Mbp)**
- **Hohe Gendichte (>90%)**
- **1 Gen = 1 ORF**
- **Gene hintereinander in Operons**
- **Ribosome Binding Sites (RBS, Shine-Dalgarno) and TK-Terminatoren.**

Statistisches Vorwissen für HMM



HMM für prokaryotisches Gen: „GeneMark“

- Betrachtet sliding window (96 nt) in DNA-Sequenz
- Berechnet Kodierungspotential für alle 6 Lese-rahmen (Inhalt)
- Lokalisiert RBS (Signal), um Startkodon genauer zu identifizieren.



Genvorhersage in Eukaryoten

GENSCAN
(Karlin and Burge 1997)

```
Predicted genes/exons:
Gn.Ex Type S .Begin ...End .Len Fr Ph I/Ac Do/T CodRg P.... Tscr..
----- ---- - ------ ------ ---- -- -- ---- ---- ----- ----- ------
1.01 Init +   1579   1663   85  0  1  114   54   131 0.741  13.24
 1.02 Intr +   2540   2635   96  0  0    1  100   134 0.698   6.38
 1.03 Intr +   3455   3588  134  0  2  101   81   136 0.999  15.07
 1.04 Intr +   4820   5042  223  1  1   85   56   432 0.998  37.93
 1.05 Intr +   5153   5350  198  0  0   73   81   371 0.999  34.74
 1.06 Intr +   5688   5889  202  1  1   53   69   378 0.979  31.27
 1.07 Intr +   6318   6426  109  0  1   62   80    20 0.843  -0.61
 1.08 Intr +   6576   6634   59  2  2  105   77    51 0.888   3.87
 1.09 Term +   6723   6792   70  0  1   63   54    98 0.785   1.61
 1.10 PlyA +   6853   6858    6                              1.05
Predicted peptide sequence(s):
>gi|GENSCAN_predicted_peptide_1|391_aa
MAMQKIFAREILDSRGNPTVEVDLHTAKGRFRAAVPSGASTGIYEALELRDGDKGRYLGK
AKFGANAILGVSLAVCKAGAAEKGVPLYRHIADLAGNPDLILPVPAFNVINGGSHAGNKL
AMQEFMILPVGASSFKEAMRIGAEVYHHLKGVIKAKYGKDATNVGDEGGFAPNILENNEA
LELLKTAIQAAGYPDKVVIGMDVAASEFYRNGKYDLDFKSPDDPARHITGEKLGELYKSF
IKNYPVVSIEDPFDQDDWATWTSFLSGVNIQIVGDDLTVTNPKRIAQAVEKKACNCLLLK
VNQIGSVTESIQACKLAQSNGWGVMVSHRSGETEDTFIADLVVGLCTGQIKTGAPCRSER
LAKYNQLMRIEEALGDKAIFAGRKFRNPKAK

Column  Description
------  -----------------------------------------------------------
Gn.Ex   gene number, exon number (for reference)
Type    Init = Initial exonIntr = Internal exon
        Term = Terminal exonSngl = Single-exon gene
        Prom = PromoterPlyA = poly-A signal
S       DNA strand (+ = input strand; - = opposite strand)
Begin   beginning of exon or signal (numbered on input strand)
End     end point of exon or signal (numbered on input strand)
Len     length of exon or signal (bp)
Fr      reading frame (a codon ending at x is in frame f = x modulo 3)
Ph      net phase of exon (exon length modulo 3)
I/Ac    initiation signal or acceptor splice site score (x 10)
Do/T    donor splice site or termination signal score (x 10)
CodRg   coding region score (x 10)
P       probability of exon (sum over all parses containing exon)
Tscr    exon score (depends on length, B/Ac, Do/T and CodRg scores)
```

Genvorhersage mit GENSCAN

- eine Beispiel

genome.mit.edu/GENSCAN.html

- max 200 kb mit mehreren Genen möglich

Predicted Exons	Annotated Exons
1579 1663	1579 1663
2540 2635	2540 2635
	2796 2854
	3016 3085
3455 3588	3455 3588
4820 5042	4820 5042
5153 5350	5153 5350
5688 5889	5688 5889
6318 6426	6318 6426
6576 6634	6576 6634
6723 6792	6723 6792

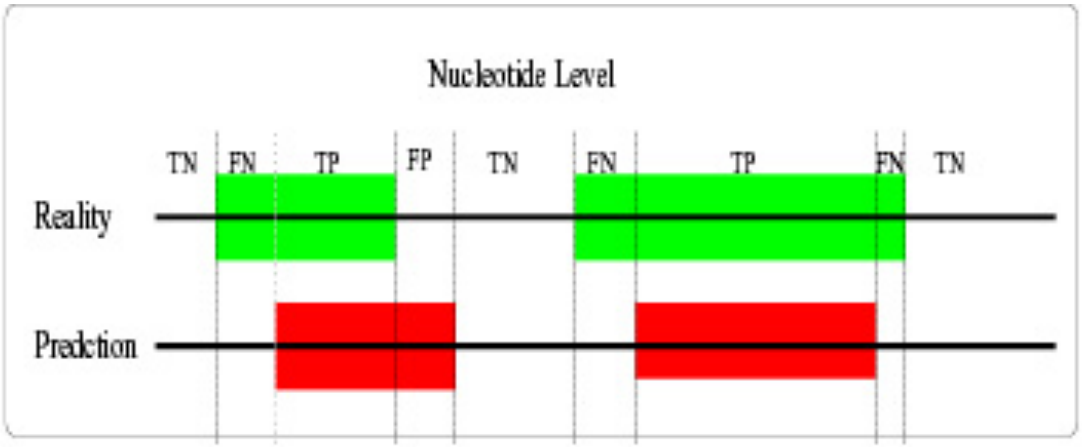
Bewertung der Genvorhersage

Sensitivität	wieviele vorhandenen Exons werden richtig erkannt? („false negatives?“)
Spezifität	wie gut wird gegenüber falschen Exons diskriminiert? („false positives?“)

Die meisten Genvorhersageprogramme erreichen mittlerweile eine Sensitivität von ca. 70 %.

Ein großes Problem bleibt auch die oft geringe Spezifität!

Bewertung der Genvorhersage

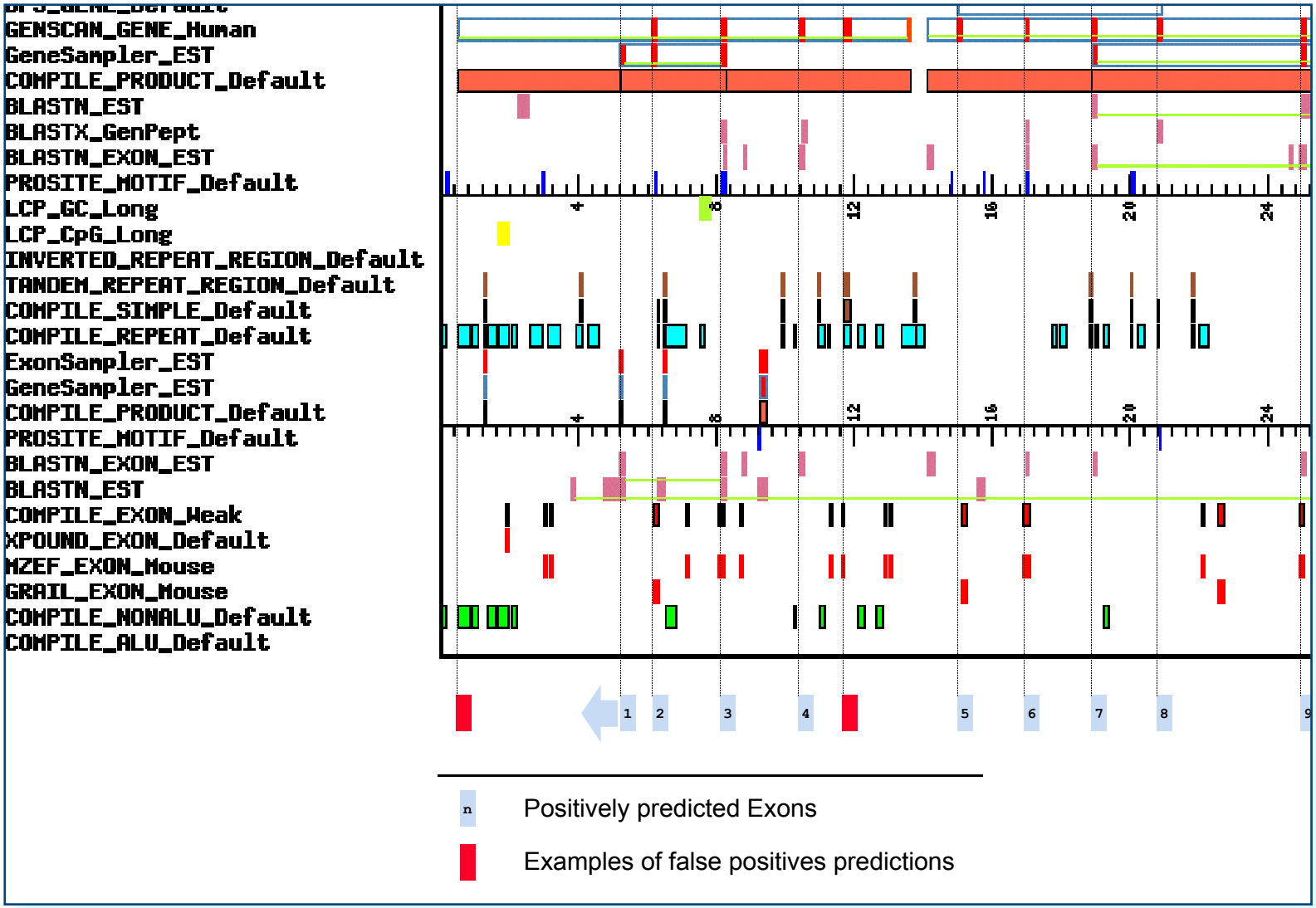


adapted from: Burset and Guigo (1996), Genomics 34, 353-367.

TN true negatives
FN false negatives
TP true positives
FP false positives

Eine mögliche Lösung:

mehrere Programme verwenden!



Genvorhersage-
Programme für eukaryotische Gene

GRAIL	nur HMM: berechnet sieben Parameter kodierender Regionen
GRAIL-EX	HMM plus BLASTN
GeneID	HMM: erst Signale berechnet, dann Kodierungspotenzial dazwischen gesucht
Genscan	HMM mit „higher order parameters“: Gendichte, typische Exon-Grösse und Anzahl, GC-Verteilung
GenomeScan	= Genscan plus BLASTX-Match zu bekannten Proteinen (Nachteil: geringere scores für Exons ohne Proteinmatch)
FGENESH	HMM , linear discriminant analysis
Genewise	HMM und parallel Proteinhomologievergleich (gut, aufwändig)
Twinscan	Genscan plus Genomvergleich
SGP-2	GeneID plus Genomvergleich

Literatur & Links

http://www.wikiwand.com/en/List_of_gene_prediction_software

Literatur:

Guigo et al., Genome Biol. 2006; 7:S2.1-31
Brent & Guigo, Curr Opin Struct Biol. 2004; 14:264-72
Yandell & Ence, Nature Reviews Genetics 2012; 13: 329

Forschungsfelder in der *ab initio*-Genvorhersage

- sehr kleine Exons
- ALT-splicing
- überlappende Gene
- Promoter-Erkennung
- 5 ‘ und 3 ‘ UTRs
- RNA-kodierende Gene
- „confidence values“ angeben

Die drei Wege zum Gen

1. Datenbanksuchen/Alignments

   „es gibt bereits passende Sequenzen in den Datenbanken“

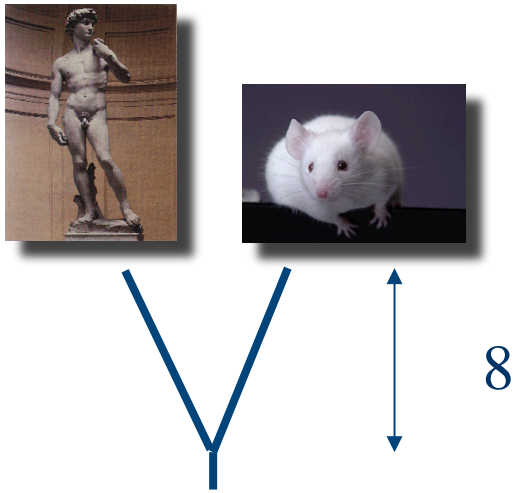
2. „*ab initio*“-Genvorhersage

   „Signale“ in der DNA zeigen: hier ist ein Gen“

3. Vergleich von verwandten Genomen
   („comparative genomics“)

   „hier ist eine evolutionär konservierte Region“

Vergleichende Genomanalyse:
die Lösung des Genvorhersageproblems



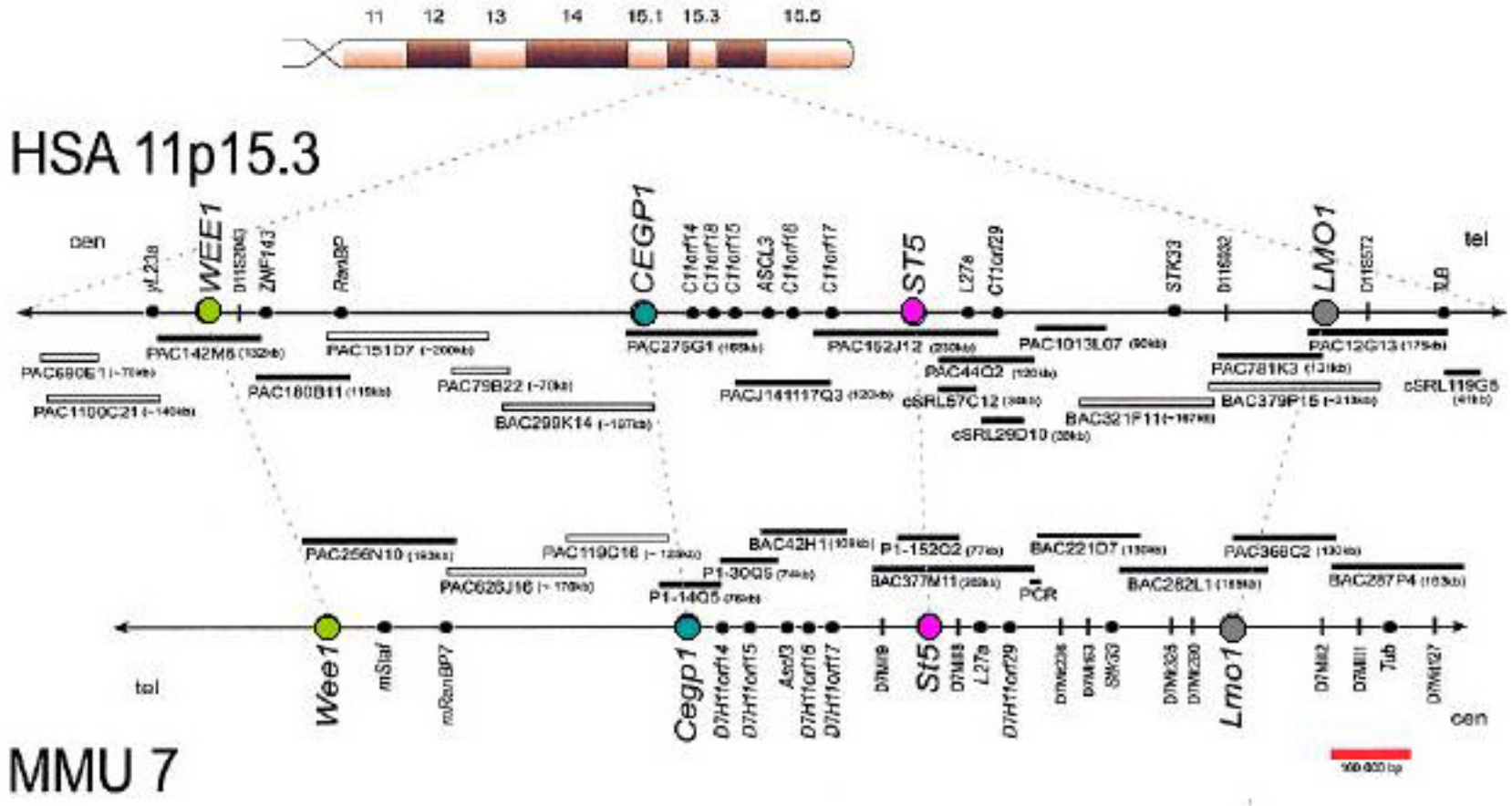
80-100 Mio. Jahre

„nur funktionell wichtige DNA-Sequenzen (Exons, regulatorische Bereiche) bleiben während der Evolution konserviert“

Unser Ziel damals (1995):

1 Mio Bp in Mensch und Maus

in hoher Qualität (<< 1 Fehler / 10 000) und annotiert



Fehlerhaftigkeit der *ab initio*-Genvorhersage

CEGP1

EXON	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	fp	%
Human	-	-	+	+	+	+	+	-	+	+	+	-	+	+	+	+	+	+	-	+	+	+	3	78
Mouse	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	4	91

C11ORF1

EXON	1	2	3	4	5	6	7	fp	%
Human	+	+	+	+	+	+	-	0	86
Mouse	+	+	+	+	+	+	+	0	100

C11ORF2

EXON	1	2	3	4	5	fp	%
Human	+	+	+	+	+	0	100
Mouse	+	+	+	+	+	0	100

OK!

ASCL1

EXON	2	fp	%
Human	+	1	100
Mouse	+	0	100

C11ORF3

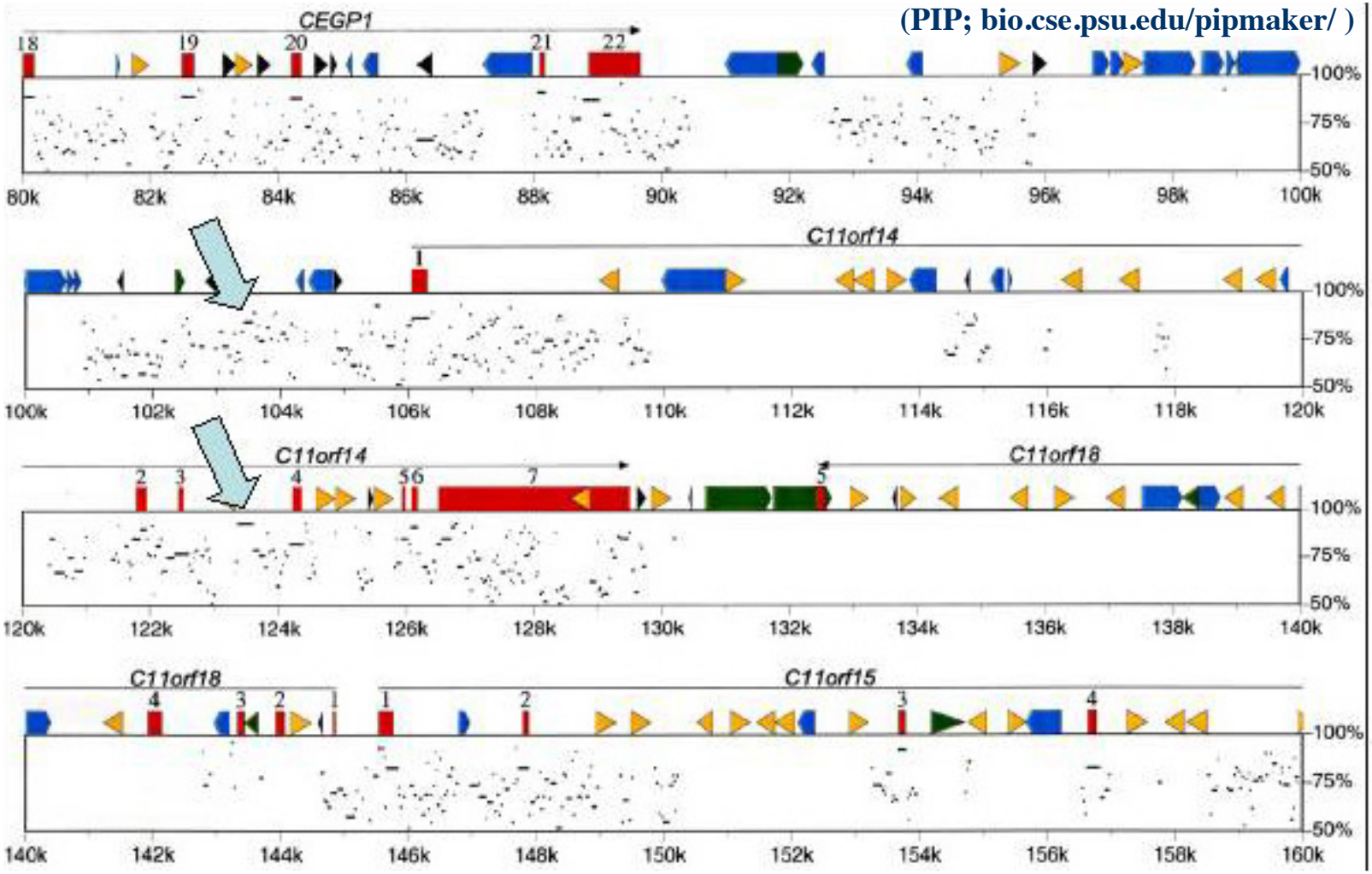
EXON	2	3	4	5	fp	%
Human	-	+	-	-	2	25
Mouse	-	+	-	-	2	25

C11ORF4

EXON	2	3	4	5	6	fp	%
Human	-	+	-	-	+	0	40
Mouse	+	+	+	-	+	0	80

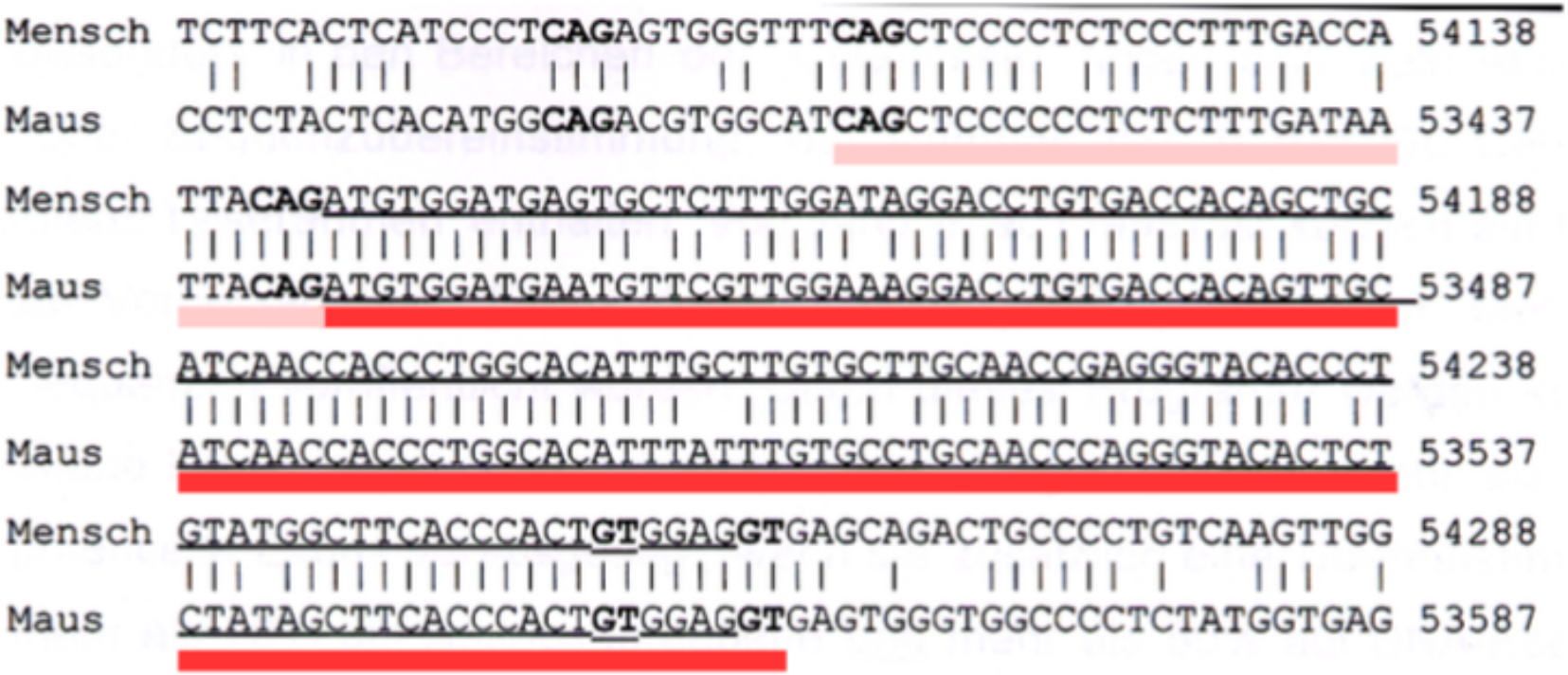
+ korrekt vorhergesagt -- falsch negativ fp falsch-positive Vorhersagen

Mensch-Maus-Sequenzvergleich mit „Percent-Identity-Plot“



Mensch-Sequenz als ‚Vorlage ‘dargestellt. Skala rechts zeigt Sequenzidentität in der Maus.

Interspezies-Vergleich verbessert Genvorhersage



Aber: die exakte Vorhersage der korrekten Spleiß-stellen ist so nicht möglich

Kombination von Genom-Vergleich und *ab initio*-Genvorhersage verbessert das Ergebnis enorm...

Bsp: Annotation von Chromosom HSA 14

Table 3 **Specificity and sensitivity of the non-expressed resources in the annotation of human chromosome 14**

	Exofish	BLAT-mouse	GENSCAN exons	FGENESH exons	Exofish + BLAT-mouse	Exofish + BLAT-mouse + GENSCAN exons + FGENESH exons
Annotation features	4,396	18,583	10,192	7,901	3,862	2,609
Specificity*	96%	31%	52%	61%	96%	99%
Sensitivity†	42%	70%	68%	63%	74%	83%

*The fraction of the annotation features of the various resources or a combination of resources that were included in the exons of the final gene annotation is indicated.
†The total number of annotated exons for all gene categories is 7,305 exons. The fraction of exons identified with the various resources or with a combination of resources is indicated.

GENSCAN, FGENESH	ab-initio Vorhersage
EXOFISH	Fugu-Genomvergleich
BLAT-mouse	Maus-Genomvergleich

Es wurde hier keine Info über cDNAs/ESTs/Proteine verwendet!!

Vergleichende Genomanalyse:

unerwartete Erkenntnisse

Ultraconserved Elements in the Human Genome

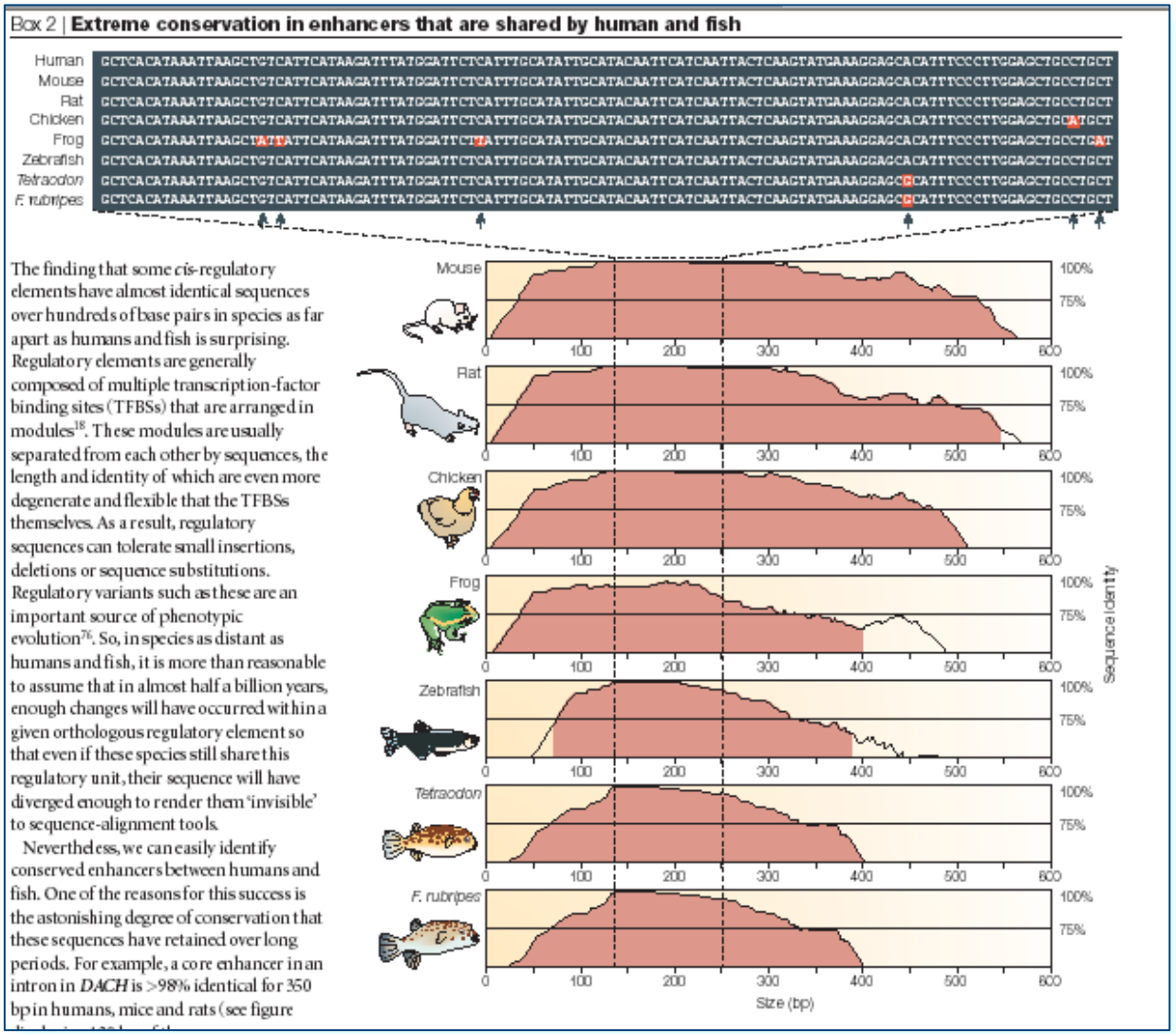
Gill Bejerano,[1*] Michael Pheasant,[3] Igor Makunin,[3] Stuart Stephen,[3] W. James Kent,[1] John S. Mattick,[3] David Haussler[2*]

There are 481 segments longer than 200 base pairs (bp) that are absolutely conserved (100% identity with no insertions or deletions) between orthologous regions of the human, rat, and mouse genomes. Nearly all of these segments are also conserved in the chicken and dog genomes, with an average of 95 and 99% identity, respectively. Many are also significantly conserved in fish. These ultraconserved elements of the human genome are most often located either overlapping exons in genes involved in RNA processing or in introns or nearby genes involved in the regulation of transcription and development. Along with more than 5000 sequences of over 100 bp that are absolutely conserved among the three sequenced mammals, these represent a class of genetic elements whose functions and evolutionary origins are yet to be determined, but which are more highly conserved between these species than are proteins and appear to be essential for the ontogeny of mammals and other vertebrates.

Vergleichende Genomanalyse: Ultraconserved Elements

Beispiel eines UCE mit möglicher Enhancer-Funktion

Warum ist UCE besser konserviert als die kodierende DNA?



a look at what is essentially an ordinary earthquake," Ellsworth says.

The mine is a challenging work environment. Moab Khotsong claims to be home to the world's deepest mine shaft: a 3000-meter descent at some 15 meters per second in a shaking cage, with only the arcs of head lamps piercing the darkness. At the drilling level, 95 floors down, it's a 10-minute ride in a rail carriage to the drill site. It can be eerily quiet, and the air carries an acrid scent of burnt rock and ammonia from recent dynamiting. "It's a smell that'd be associated with Hades," says Tullis Onstott, a geomicrobiologist at Princeton University who has also joined the project.

Onstott's goal is to learn whether earthquakes can favor the microbes that live in deep rock. In an experiment he's been trying to do for a decade, ever since it failed at a San Andreas drill site, he will install an automated sampling system, triggered by seismicity in the borehole at the fault. Past work has shown that an earthquake can release a pulse of hydrogen gas, which might have been trapped in the rock or generated by chemical reactions caused by the fracturing. Onstott hopes, as more aftershocks strike, to discover a cascade of microbial populations feeding off this chemical energy. If so, similar faults could become targets for a search for life on Mars, which might influence the selection of a landing site for NASA's Mars 2020 rover mission, Onstott says.

Mine operators might also get what they are seeking: a clearer picture of how human activity can trigger quakes. In the United States, the underground disposal of wastewater from oil and gas drilling is known to trigger quakes by boosting pressure in the rock pores. Pore pressure is unlikely to be a factor in South Africa, says Ze'ev Reches, a geophysicist at the University of Oklahoma in Norman. But he notes that the Orkney quake, like some of the largest induced quakes in Oklahoma, occurred deep in basement rock along a previously unmapped fault. (Reches is planning to drill into a fault in Oklahoma, but he'll have to go 4.5 kilometers down.)

Plenty can still go wrong for DSeis. In May, operators shuttered another mine that the project planned to use, citing an increased risk of falling rocks. But Ogasawara says they have plenty of candidate earthquake faults for their campaign, including small, shallow ones accessible from nearby mines, which may allow them to target the very heart of an earthquake: the hypocenter, the exact place where a fault first starts to give way. "That is what we want to see," Ogasawara says. "We're expecting to show what a hypocenter looks like." ■

GENOMICS

Mysterious unchanging DNA finds a purpose in life

Ultraconserved elements have developmental roles

By **Elizabeth Pennisi**,
in Cold Spring Harbor, New York

"First, thank you. Many of us have been waiting for this." That's how Eric Lander, director of the Broad Institute in Cambridge, Massachusetts, greeted genome scientist Diane Dickel earlier this month after she finished a talk on mysterious genetic sequences known as ultraconserved elements (UCEs). Speaking at the Biology of Genomes meeting here, Dickel had presented what may be the first compelling evidence of what these much-discussed stretches of DNA do. Most biologists are convinced they are important, as they have barely changed over tens of millions of years. But clues to their function had been scarce until Dickel's findings, which suggest they may play key roles in animal development.

> ***"The bottom line: These things are important in biology."***
>
> **Diane Dickel**, Lawrence Berkeley National Laboratory

Computational scientists first identified these anomalous stretches of DNA 13 years ago, by comparing the genomes of diverse mammals (*Science*, 28 May 2004, p. 1321). The University of California (UC), Santa Cruz, researchers found 481 genome segments, 200 to more than 700 bases long, that were 100% identical in humans, rats, and mice. These same segments were at least 95% identical in chickens and dogs, and many were also strongly conserved in fish. Many UCEs were also located outside genes, in regions of chromosomes many researchers considered nonfunctional. Why would this "junk" be conserved? "That was quite baffling," recalls Kelly Frazer, a genome scientist at UC San Diego.

Apparently off limits to evolutionary tinkering, UCEs presumably had some crucial function. Change a UCE, and an animal should die. Or so researchers thought. Three years later, a team from Lawrence Berkeley National Laboratory (LBNL) in Berkeley, California, created four genetically engineered mouse strains, each with a different UCE knocked out. The results were startling: The mice seemed to be just fine.

A decade later, Dickel, part of the same LBNL lab group that couldn't originally find a purpose for UCEs, tried again, this time with the gene-editing technique called CRISPR. She wondered whether some UCEs are redundant, which would allow intact elements to fill in for deleted ones, so she used CRISPR to delete more than one UCE at a time.

She focused on four UCEs near the *ARX* gene, which codes for a transcription factor that regulates genes for proper brain and sexual development. One of those had been deleted by the original LBNL group, and all four elements are thought to be enhancers, stretches of regulatory DNA that drive gene activity, two in the upper forebrain of a developing mouse and two in the lower forebrain.

At first, knocking out these elements, whether singly or in pairs, seemed to have the mice unaffected. But when Dickel and a developmental neuroscientist took a close look at the brains of the mice, they detected abnormal numbers of neurons or a shrunken hippocampus—effects visible in both the double and, to a lesser degree, the single knockouts. "This was very satisfying in that it resolved a long-standing observation that really bothered us," Dickel says.

Although the brain differences may be too small to be significant for well-fed, coddled lab mice, they might be catastrophic for mice in the wild, where they struggle to find food and avoid predation, Dickel notes. And any evidence of a role for UCEs is welcome. As Harvard University geneticist Chao-ting Wu puts it: "It's really lovely that there is some function."

Dickel now plans to use CRISPR to surgically knock out individual bases in the ultraconserved *ARX* regions. Other researchers have suggested that UCEs in enhancers vary so little because they need to bind to multiple proteins to regulate gene activity—and the loss of single base would disable them. Dickel isn't convinced yet that they're right, but she's happy to have finally provided some purpose for UCEs. "The bottom line: These things are important in biology." ■