



Genvorhersage bei Eukaryoten

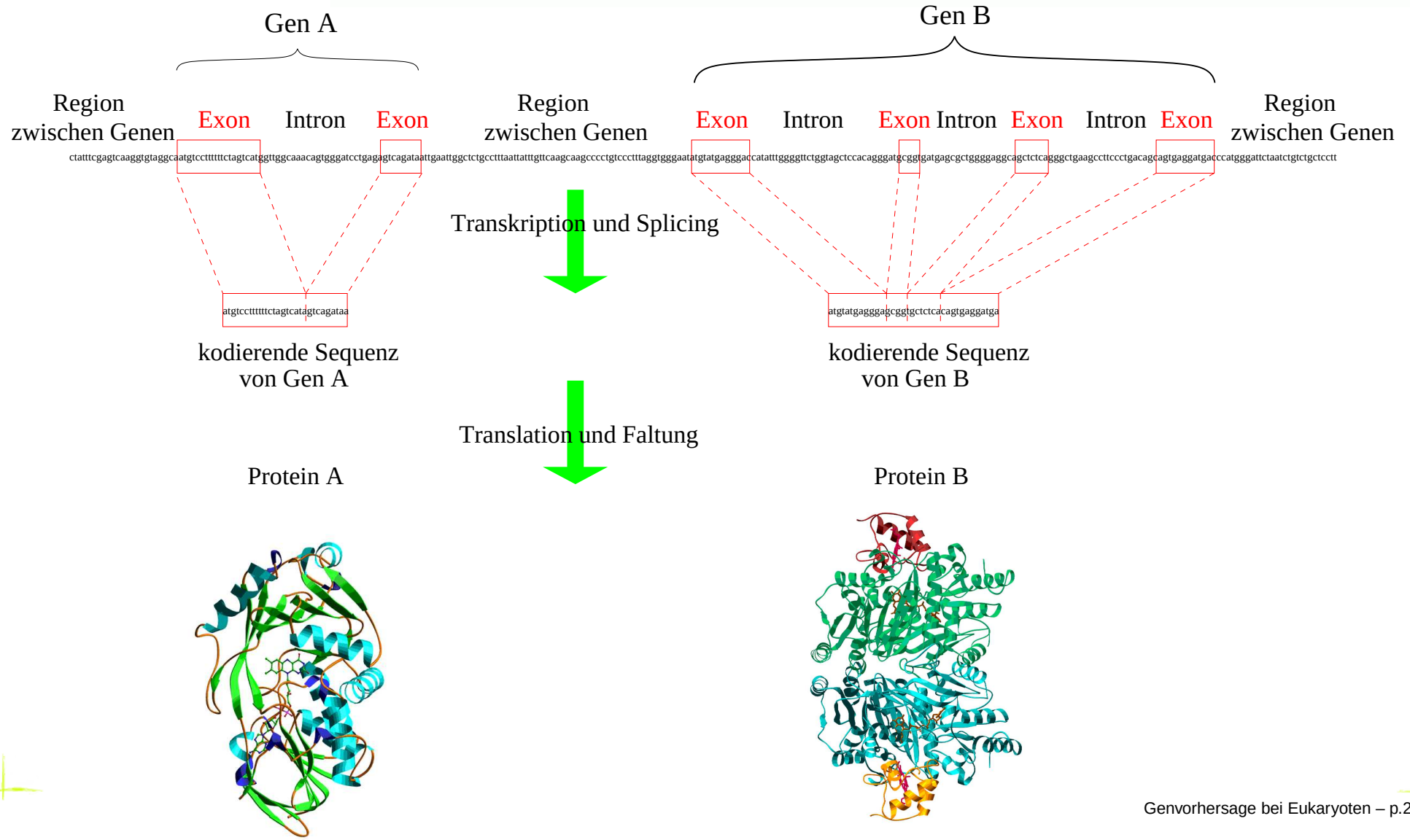
Mario Stanke

mstanke@gwdg.de

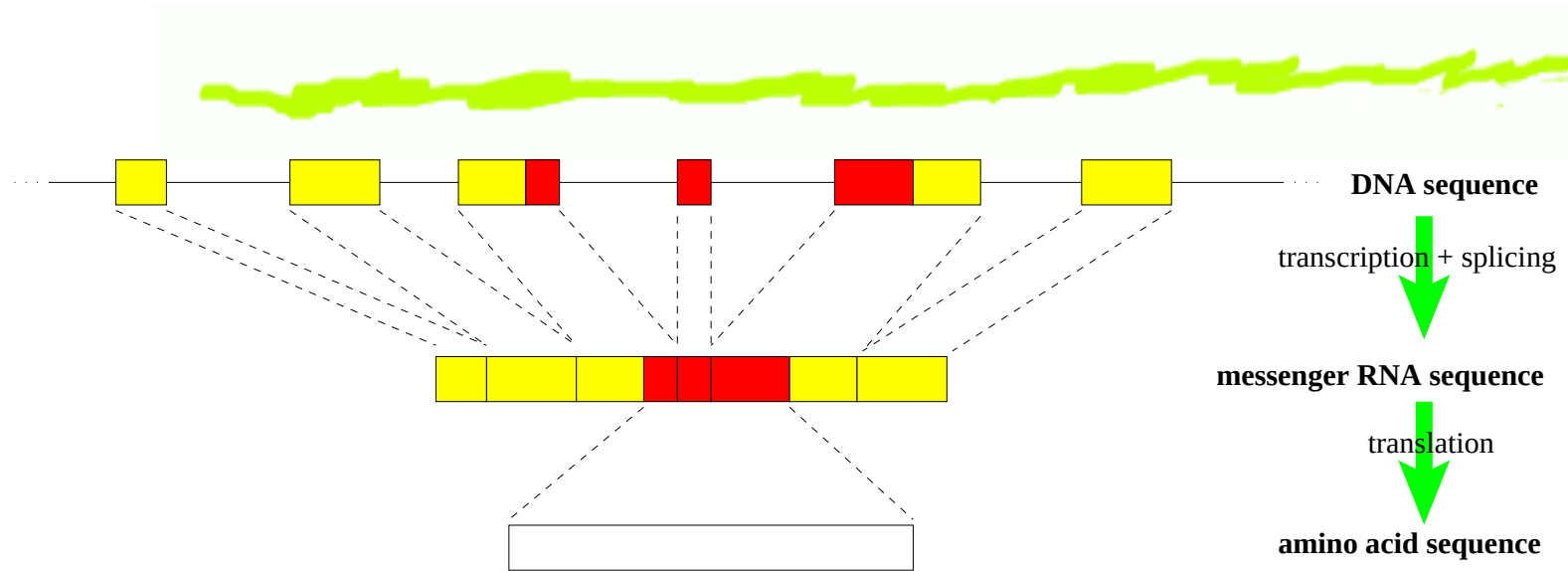
Institut für Mikrobiologie und Genetik

Abteilung Bioinformatik

Was wird vorhergesagt?



Was wird vorhergesagt?



Legend:

— not transcribed or transcribed and spliced out transcribed, not spliced out but not translated translated

Beispiel eines menschlichen Gens

cctcacctctgagaaaacctctttgccaccaataccatgaagctctgcgtgactgtcctgtctctc
ctcgtgctagtagctgccttctgctctcttagcactctcagcaccaagtaagtctacttttgcagct
gctatttctgagtcaaggtgtaggcagagtccttttttctagtcattggctggcaaacagtgggatct
ggggatgggacaaaaggcagctaggaagattgccatgtagtctgctgctaaatgtagagtctagta
gatattcagtaacattcaagttcctattttcttaagaattagcaaccagcagaggaaaacgatggg
ctggaagtcagactgttgaattggctctgcctttaattatttgttcaagcaagccctgtccctct
ctgtgccttggtttcccatctgtcataatgaaggagtgcatgtgttctgagactgaatccagtt
ccaatcttctagattttctttctcgttcttctctgaagatccactattcagaataagactcctgtct
atgttaggtgggaatggatacaagggaccataatttgggggttctggtagctccacagggatgctcaa
tgaagatgcaaaattagaagtcaaaataaacagctcccatgggcagtggtgatctcaccctggcct
ttccttttcagtgggctcagaccctcccaccgcctgctgcttttcttacaccgcgaggaagcttcct
cgcaactttgtggtagattactatgagaccagcagcctctgctcccagccagctgtgggtgtgagta
tcaacccttggtgcccctgggaggcaagggtgagggtggtatttttaaagggggcctgttttgggg
agggggtgatgagcgtggggaggcagctctcagggtgaagccttccctgacagcagtgagggtca
cagggtcatgaactcacttttcaagtgtgaaggcggctgagtggcagccgagacagaagggggttc
ctggggagggaagttattcagaggacagggaagcaggggaaggcagacagggtcccatgagatatgga
ccaattccttaaaccatgctagaaaaacatgtggaaaagtcactaccagggtggcagggaatgggg
caatctattcatactgattgcaatgccactgggttcctaatctgggcaaccctggggccacagc
taaattccagtgagtggaagttacagggtgctgcttccagtgtgctcagaggaaggatcccatcca
ccagagctgccccacatggaccatggtcaggcagaggaagatgcctaccacaggcaagggaataaag
ccagatgacctcaaagggtcccatgggattctaatctgtctgctccttgttctacagattccaaacc
aaaagaggcaagcaagtctgctgctgaccccagtgagtcctgggtccaggagtacgtgtatgacctg
gaactgaactgagctgctcagagacaggaagtcttc

Beispiel eines menschlichen Gens

cctcacctctgagaaaacctctttgccaccaataccatgaagctctgcgtgactgtcctgtctctc
ctcgtgctagtagctgccttctgctctctagcactctcagcaccaagtaagtctacttttgcagct
gctatttgcagtcagggtgtaggcagagtccttttttctagtcattggctggcaaacagtgggatct
ggggatgggacaaaaggcagctaggaagattgccatgtagtctgctgctaaatgtagagtctagta
gatattcagtaacattcaagttcctattttcttaagaattagcaaccagcagaggaaaacgatggg
ctggaagtcagactgttgaattggctctgcctttaattatttgttcaagcaagcccctgtccctct
ctgtgccttggtttcccatctgtcataatgaaggagtgcatgtgttctgagactgaatccagtt
ccaatcttctagattttctttctcgttcttctctgaagatccactattcagaataagactcctgtct
atgttaggtgggaatggatacaaggaccataatttgggggttctggtagctccacagggatgctcaa
tgaagatgcaaaaattagaagtcaaaataaacagctcccatgggcagtggtgatctcaccctggcct
ttccttttcagtgggctcagaccctcccaccgcctgctgctttttcttacaccgcgaggaagcttct
cgcaactttgtggtagattactatgagaccagcagcctctgctcccagccagctgtggtgtgagta
tcaacccttggtgcccctgggaggcaagggtgagggctggatttttaaagggggcctgttttgggg
agggggtgatgagcgtggggaggcagctctcagggctgaagccttccctgacagcagtgagggtca
caggtcatgaactcacttttcaagtgtgaaggcggctgagtggcagccgagacagaagggggttc
ctggggaggaagttattcagaggacagggaagcaggggaaggcagacagggtcccatgagatatgga
ccaattccttaaaccatgctagaaaaacatgtggaagagtcactaccaggctggcagggaatgggg
caatctattcatactgattgcaatgccactgggttcctaatctgggcaaccctggggcccacagc
taaattccagtgagtggaagttacaggagagtcctgcttccagtgctgctcagaggaaggatcccatcca
ccagagctgccccacatggaccatgggtcaggcagaggaagatgcctaccacaggcaagggaataaag
ccagatgacctcaaagggtcccatgggattctaatctgtctgctccttgttctacagattccaaacc
aaaagaggcaagcaagctctgcgctgacccagtgagtcctgggtccaggagtacgtgtatgacctg
gaactgaactgaagctgctcagagacaggaagtcctc

Warum Genvorhersage mit dem Computer?

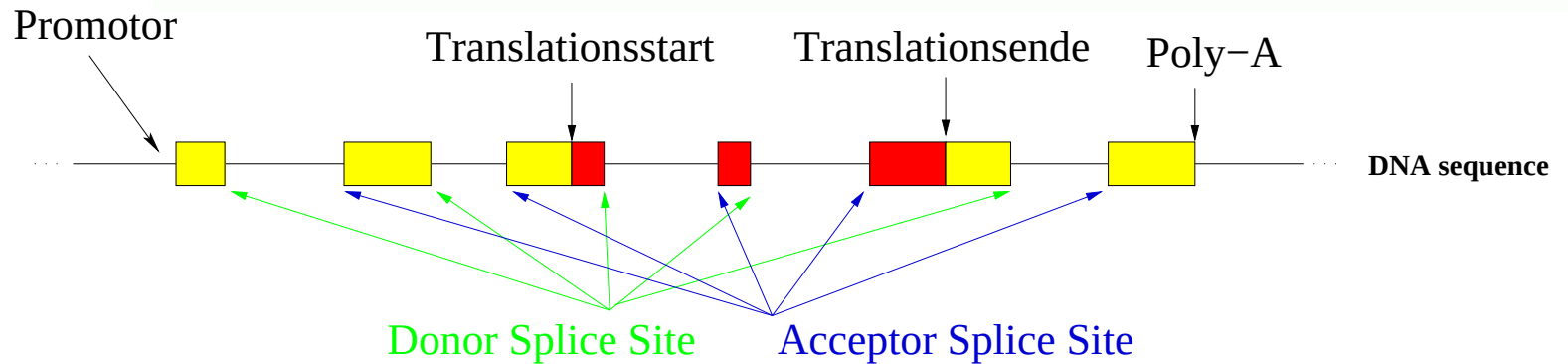
- ⑥ geht nicht anders, weil sehr viele Sequenzdaten:
39 eukaryotische Genome sequenziert
mindestens 547 zusätzliche eukaryotische Genome
angefangen
- ⑥ experimentelle Lokalisierung von Genen beruht auf
exprimierten Sequenzen, aber Gen kann
'abgeschaltet' sein in Probe
- ⑥ Vorhersagen sind zwar unsicher, können aber helfen,
die tatsächliche Genstruktur zu finden: gezielte,
experimentelle Verifizierung vorhergesagter Gene

Methoden der Genvorhersage



- ⑥ Signalsensoren
 - ⑥ Inhaltssensoren
 - ⑥ Open Reading Frames
 - ⑥ Inter-Spezies DNA Vergleiche
 - ⑥ cDNA-Alignment
 - ⑥ EST-Alignments
 - ⑥ Protein-Homologie
- } ab initio Methoden
- } extrinsische Methoden

Signalsensoren



Signale sind kurze Abschnitte auf der DNA, die die Translation oder Transkription steuern.

Signalsensoren

- ⑥ Promotor: Markiert (falls vorhanden) den Beginn der Transkription.
- ⑥ Splice Sites: 5' (Donor) und 3' (Akzeptor) Ende des Introns, das herausgespleißt wird
- ⑥ Translationsstart: Enthält das StartKodon (meistens ATG) und markiert den Beginn der Translation.
- ⑥ Translationsende: Enthält das Stoppcodon (TGA, TAA oder TAG) und markiert das Ende der Translation.
- ⑥ Poly-A Signal: Leitet das Ende der Transkription ein.

Donor Splice Site



Diese Signale enthalten zwar **typische** Sequenzmotive, diese Motive sind aber nicht **charakteristisch**: Die Motive treten auch auf, wo keine Signale sind.

Beispiel Donor Splice Site. (Fast) jedes Intron beginnt mit den beiden Nukleotiden **gt**, aber das reicht nicht um Donor Splice Sites zu finden.

Donor Splice Site

cctcacctctgagaaaacctctttgccaccaataccatgaagctctgctgtgactgtcctgtctctc
ctcgtgctagtagctgccttctgctctcttagcactctcagcaccaagttaaagtctacttttgcagct
gctatttctgagtcaagggtgtaggcagagtccttttttctagtcatggctggcaaacagtgggatct
ggggatgggacaaaaggcagctaggaagattgccatgtagtctgctgctaaatgtagagtctagta
gatattcagtaaacattcaaggttcctattttcttaagaattagcaaccagcagaggaaaacgatggg
ctggaaagtcagactgttgaattggctctgctttaaattatttgttcaagcaagcccctgtccctct
ctgtgccttggtttcccatctgtcataatgaagggaagtcgatgtgttctgagactgaatccaagt
ccaatcttctagattttctttctcgttcttctctgaagatccactattcagaataagactcctgctc
atgttaggtgggaatggatacaagggaccataatttggggttctggtagctccacagggatgctcaa
tgaagatgcaaaattagaaagtcaaaataaacagctcccatgggcaagtgttgatctcacccctggcct
ttccttttcagtgggctcagaccctcccaccgcctgctgcttttcttacaccgcgaggaagcttcct
cgcaactttgtggtagattactatgagaccagcagcctctgctcccagccagctgtgggtgtgagta
tcaacccttggtgcccctgggaggcaagggtgagggctggatttttaaagggggcctgttttgggg
aggggtgatgagcgtggggaggcagctctcagggctgaagccttccctgacagcagtgaggtca
caggtcatgaactcacttttcaaagtgctgaaggcggctgagtggtgagccgagacagaagggggttc
ctggggaggaaattattcagaggacagggaagcaggggaaggcagacaggtcccatgagatatgga
ccaattccttaaaccatgctagaaaaacatgtggaaaagtactaccaggctggcagggaatgggg
caatctattcatactgattgcaatgccactgttcctaattctgggcaaccctggggccacagc
taaattccaagtgaagtgaagtgttacagggaagtcgttccaagtgtgctgctcaggaaggatcccatcca
ccagagctgccccacatggaccatgtcaggcagaggaagatgcctaccacaggcaagggaataaag
ccagatgacctcaaaggtcccatgggattctaatctgtctgctccttgttctacagattccaaacc
aaaagaggcaagcaaagtcctgctgaccccaagtgaagtcctgggtccaggagtaggtatgacctg
gaactgaactgagctgctcagagacaggaaagtcct

Donor Splice Site

Exon

```
...agagcaaggtacgc...
...tctttcatgtgagt...
...actgtcaggtatgt...
...gacaaaaggtacgt...
...tccaaaaggcaggg...
...tttcctaggtaacg...
...agaagatggtagga...
...tcctttgggtgagt...
...gatcctgggtcagt...
...cacgctgggtacgc...
...atcacctcgtgagt...
...cttccagggtgaga...
...acctcacggtgaga...
...aacaggaggtacca...
...gtcggaccgtgagt...
...aaactgcggtgagt...
...ccagggaaggtaggg...
...gcactgtggtgagc...
...gaggacaggtgagc...
...cgagggcggtgagc...
...cctgtcaggtgagt...
...acgtggaggtgagg...
...atcagtatgtgagt...
...tacatcgggtcagt...
...gccaccaggtaggg...
...gatctgtggtgaga...
...tctggccggtttgt...
...cttccaaggtaggg...
...gcgccaaaggttggc...
```

Intron

Donor Splice Site



Ein Pictogramm der Region -8 bis 6 bezogen auf die Grenze zwischen Exon und Intron. Die Größe der Buchstaben der Basen ist proportional zur relativen Häufigkeit der Base an dieser Stelle.

Inhaltssensoren

Kodierende Sequenzen (kodierende Teile von Exons) und *nicht-kodierende* Sequenzen (Introns, zwischengenische Region) haben auch typischerweise verschiedene **Basenzusammensetzungen**. Zum Beispiel:

- ⑥ kodierend: hoher Gehalt an g und c
- ⑥ nicht-kodierend: hoher Gehalt an a und t

Inhaltssensoren

cctcacctctgagaaaacctctttgccaccaataccatgaagctctgcgtagactgtcctgtctctc
ctcgtgctagtagctgccttctgctctctagcactctcagcaccaagtaagtctacttttgcagct
gctatttgcagtcagggtgtaggcagagtccttttttctagtcattggctggcaaacagtgggatct
ggggatgggacaaaaggcagctaggaagattgccatgtagctgtgctgctaaatgtagagcttagta
gatattcagtaacattcaagttcctattttcttaagaattagcaaccagcagaggaaaacgatggg
ctggaagtcagactgttgaattggctctgcctttaattattttgttcaagcaagcccctgtccctct
ctgtgccttggtttcccatctgtcataatgaaggagtgcatgtgttctgagactgaatccagtt
ccaatcttctagattttctttctcgttcttctctgaagatccactattcagaataagactcctgtct
atgttaggtgggaatggatacaagggaccataatttgggggttctggtagctccacagggatgctcaa
tgaagatgcaaaattagaagtcaaaataaacagctcccatgggcagtggtgatctcaccctggcct
ttccttttcagtgggctcagaccctcccaccgctgctgctttttcttacaccgcgaggaagcttcct
cgcaactttgtggtagattactatgagaccagcagcctctgctcccagccagctgtggtgtgagta
tcaacccttggtgcccctgggaggcaagggtgagggtggtatttttaaagggggcctgttttgggg
agggggtgatgagcgtggggaggcagctctcagggtgaagccttccctgacagcagtgagggtca
cagggtcatgaactcacttttcaagtgtgaaggcggctgagtggcagccgagacagaagggggttc
ctggggaggaagttattcagaggacagggaagcaggggaaggcagacagggtcccatgagatatgga
ccaattccttaaaccatgctagaaaaacatgtggaagagtcactaccagggtggcagggaatggg
caatctattcatactgattgcaatgccactgggttcctaatctgggcaaccctggggccacagc
taaattccagtgagtggaagttacagggtgctgcttccagtgtgctcagaggaaggatcccatcca
ccagagctgccccacatggaccatggtcaggcagaggaagatgcctaccacaggcaagggaataag
ccagatgacctcaaagggtcccatgggattctaatctgtctgctccttgttctacagattccaaacc
aaaagaggcaagcaagtctgcgctgacccagtgagtcctgggtccaggagtacgtgtatgacctg
gaactgaactgagctgctcagagacaggaagtcttc

Inhaltssensoren

Die Häufigkeiten von **Dinukleotiden** können noch besser der Unterscheidung von kodierenden und nicht-kodierenden Sequenzen dienen als die Häufigkeiten der einzelnen Basen.

Beispiel: Das Dinukleotid **at** kommt häufiger in nichtkodierenden als in kodierenden Sequenzen vor.

Inhaltssensoren

cctcacctctgagaaaacctctttgccaccaataccatgaagctctgcgtgactgtcctgtctctc
ctcgtgctagtagctgccttctgctctctagcactctcagcaccaagtaagtctacttttgcagct
gctatttcgagtcaaggtgtaggcagagtccttttttctagtcattggctggcaaacagtgggact
ggggatgggacaaaaggcagctaggaagattgccatgtagtctgctgctaaatgtagagtctagta
gatattcagtaacattcaagttcctattttcttaagaattagcaaccagcagaggaaaacgatggg
ctggaagtcagactgttgaattggctctgcctttaattattttgttcaagcaagcccctgtccctct
ctgtgccttggtttcccccattctgtcatatgaaggagtgcatgtgttctgagactgaatccagtt
ccaattcttctagattttctttctcgttcttctctgaagattccactattcagaataagactcctgctc
atgttaggtgggaatggatacaaggaccatatttgggggttctggtagctccacagggatgctca
tgaagatgcaaaattagaagtcaaaataaacagctcccattgggcagtgttgatctcaccctggcct
ttccttttcagtgggctcagaccctcccaccgcctgctgctttttcttacaccgcgaggaagcttcct
cgcaactttgtggtagattactatgagaccagcagcctctgctcccagccagctgtggtgtgagta
tcaacccttggtgcccctgggaggcaaggggtgagggctggatttttaaagggggcctgttttgggg
aggggggtgatgagcgtggggaggcagctctcagggctgaagccttccctgacagcagtgaggtca
caggtcatgaactcacttttcaagtgtgaaggcggctgagtggcagccgagacagaaggggggttc
ctggggaggaagttattcagaggacaggggaagcaggggaaggcagacaggtcccattgagatatgga
ccaattccttaaaccattgctagaaaaaacattgtgaaaagtcactaccaggctggcagggaatgggg
caattctattcatactgattgcaatgccactgggttcctaattctgggcaaccctggggccacagc
taaattccagtgagtgaagttacagggagtcctgcttccagtgctgctcgaggaaggattcccattcca
ccagagctgccccacattggaccattggtcaggcagaggaagatgcctaccacaggcaagggaataaag
ccagatgacctcaaagggtcccattgggattctaatctgtctgctccttggttctacagattccaaacc
aaaagaggcaagcaagtctgcgctgacccagtgagtcctgggtccaggagtacgtgtatgacctg
gaactgaactgagctgctcagagacaggaagtcttc

Markow Ketten

Sei für zwei Basen x und y

$p_{\text{kod}}(x|y) =$ relative Häufigkeit von x nach y
in **kodierenden** Sequenzen

und

$p_{\text{nkod}}(x|y) =$ relative Häufigkeit von x nach y
in **nicht-kodierenden** Sequenzen .

Markow Ketten

Für eine Stück Sequenz $s = x_1, x_2, \dots, x_n$, von dem man nicht weiß, ob es kodierend ist, sei

$$P_{\text{kod}}(s) := p_{\text{kod}}(x_2|x_1) \cdot p_{\text{kod}}(x_3|x_2) \cdot \dots \cdot p_{\text{kod}}(x_n|x_{n-1})$$

und entsprechend sei $P_{\text{nkod}}(s)$ definiert. Wir vermuten, daß s kodierend bzw. nicht-kodierend ist, je nachdem ob

$$P_{\text{kod}}(s) > P_{\text{nkod}}(s)$$

oder nicht.

extrinsische Information

Von “außen” kommende Hinweise über die Genstruktur:

- ⑥ cDNA-Alignment
- ⑥ Expressed Sequence Tags (ESTs)
- ⑥ Ähnlichkeiten zu bekannten Proteinen
- ⑥ Vergleich der DNA von 2 oder mehr Arten

Ähnlichkeiten zu bekannten Proteinen

Benutzen lokale Ähnlichkeit zwischen übersetzter DNA Eingabesequenz und einer Aminosäuresequenz aus einer Protein-Datenbank um Hinweise auf kodierende Regionen zu erhalten.

Beispiel:

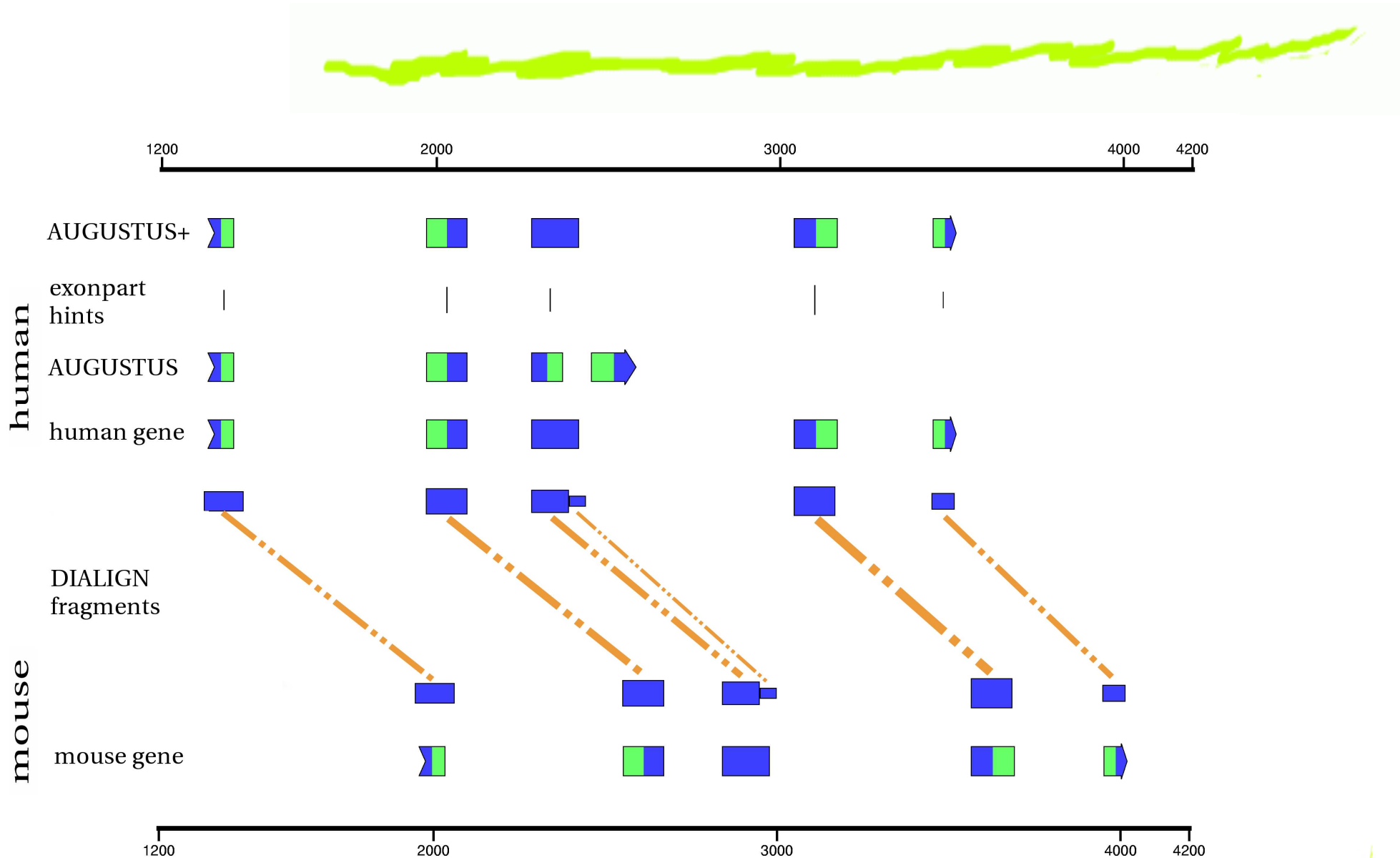
menschliche DNA:	gcc atg tcg tcc ggc atc cat gta gcg ctg gtg act gga ggc aac aag ggc atc ggc ...
übersetzte Seq.:	H V A L V T G G N K G I G ...
A. thaliana Homolog:	N V A V V T G S N R G I G ...

Vergleiche der DNA zwischen Arten



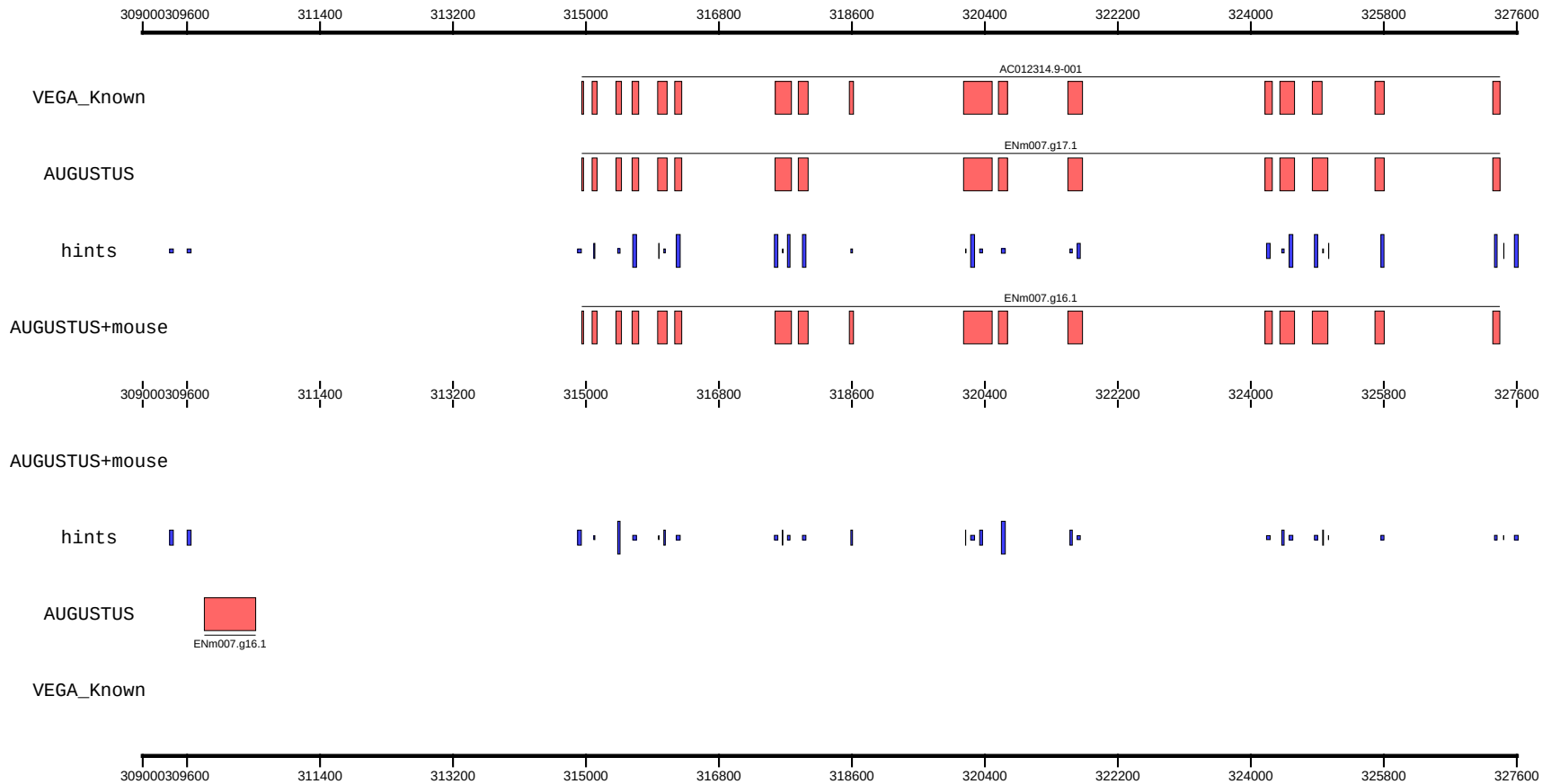
Betrachten die DNA-Sequenzen zweier Arten, die für ein ähnliches Protein kodieren. Abschnitte der Sequenz, die eine **Funktion** haben (insbesondere kodierende Bereiche) sind tendenziell **mehr konserviert**.

Vergleiche der DNA zwischen Arten



Vergleiche der DNA zwischen Arten

ENm007



This plot has been obtained using gff2ps. The most recent version of gff2ps is freely available at "<http://www.limn.es/software/gff2ps.html>". Copyright © 1999 by Josep F. Abril & Roderic Guigo

Screenshot AUGUSTUS Web Interface

The screenshot shows a web browser window titled "Augustus: job submission - Konqueror". The address bar displays "http://augustus.gobics.de/submission". The main content area has a dark blue header with the text "Augustus [job submission]". Below this, there is a large text input field with the placeholder text "Paste your sequence here:". Below the text field, there is a section for file upload with the text "or upload a file in (multiple) FastA format" and a "Durchsuchen..." button. Further down, there are radio buttons for "Organism:" with "Human" selected and "Drosophila" as an option. Below that, there are radio buttons for "Report genes on:" with "both strands" selected, and "forward strand only" and "reverse strand only" as other options. At the bottom, there are two buttons: "Reset all input" and "Run AUGUSTUS".

Augustus: job submission - Konqueror

Dokument Bearbeiten Ansicht Gehe zu Lesezeichen Extras Einstellungen Fenster Hilfe

Adresse: <http://augustus.gobics.de/submission>

Augustus [job submission]

Paste your sequence here:

or upload a file in (multiple) FastA format

Durchsuchen...

Organism: ☒ Human ☐ Drosophila

Report genes on: ☒ both strands ☐ forward strand only ☐ reverse strand only

Reset all input Run AUGUSTUS

Probleme der Genvorhersage



- ⑥ alternatives Splicing
- ⑥ TIS und Promotorvorhersage weniger zuverlässig
- ⑥ Ausnahmen, z.B.
 - △ sich überlappende Gene
 - △ nicht-kanonische Spleißstellen
 - △ (programmierte) Frameshifts
 - △ abweichende genetische Codes

Probleme der Genvorhersage



- ⑥ alternatives Splicing
- ⑥ TIS und Promotorvorhersage weniger zuverlässig
- ⑥ Ausnahmen, z.B.
 - △ sich überlappende Gene
 - △ nicht-kanonische Spleißstellen
 - △ (programmierte) Frameshifts
 - △ abweichende genetische Codes

Probleme der Genvorhersage



- ⑥ alternatives Splicing
- ⑥ TIS und Promotorvorhersage weniger zuverlässig
- ⑥ **Ausnahmen**, z.B.
 - △ sich überlappende Gene
 - △ nicht-kanonische Spleißstellen
 - △ (programmierte) Frameshifts
 - △ abweichende genetische Codes

Probleme der Genvorhersage



- ⑥ alternatives Splicing
- ⑥ TIS und Promotorvorhersage weniger zuverlässig
- ⑥ Ausnahmen, z.B.
 - △ sich überlappende Gene
 - △ nicht-kanonische Spleißstellen
 - △ (programmierte) Frameshifts
 - △ abweichende genetische Codes

Verlässlichkeit der Genvorhersage

hängt stark ab vom Organismus und von der Menge der vorhandenen ESTs/cDNAs

exprimierte Sequenzdaten > vergleichende Genvorhersage > ab initio
ab initio, nur Humangenom:

- ⑥ Etwa 40% der Exons können nicht korrekt vorhergesagt werden.
- ⑥ Etwa 40% der vorhergesagten Exons sind nicht korrekt.
- ⑥ Höchstens ein Viertel der Gene werden komplett korrekt vorhergesagt.