



Gene Prediction

Mario Stanke

mstanke@gwdg.de

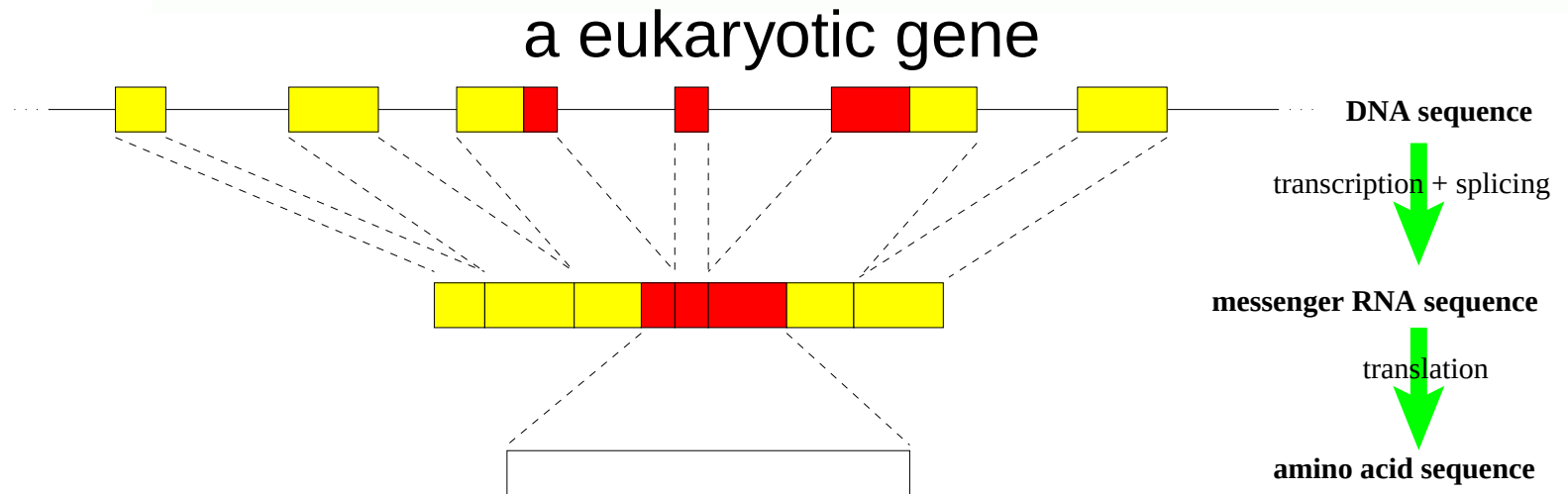
Institut für Mikrobiologie und Genetik

Abteilung Bioinformatik

Why Predict Genes with a Computer?

- ⑥ tons of data
 - △ 39/250 eukaryotic/prokaryotic genomes sequenced
 - △ 547/804 additional sequencing projects started
- ⑥ no experimental evidence available: gene not expressed in sample
- ⑥ selective experimental verification of predictions

What is Predicted?



Legend:



Example of a Human Gene

cctcacctctgagaaaacctctttgccaccaataccatgaagctctgcgtagactgtcctgtctctc
ctcgtgctagtagctgccttctgctctcttagcactctcagcaccaagtaagtctacttttgcagct
gctatttctgagtcaaggtgtaggcagagtccttttttctagtcattggctggcaaacagtgggatct
ggggatgggacaaaaggcagctaggaagattgccatgtagtctgctgctaaatgtagagtctagta
gatattcagtaacattcaagttcctattttcttaagaattagcaaccagcagaggaaaacgatggg
ctggaagtcagactgttgaattggctctgacctttaattatttgttcaagcaagccccctgtccctct
ctgtgccttggtttcccatctgtcataatgaaggagtgcatgtgttctgagactgaatccagtt
ccaatcttctagattttctttctcgttcttctctgaagatccactattcagaataagactcctgtctc
atgttaggtgggaatggatacaagggaccataatttgggggttctggtagctccacagggatgctcaa
tgaagatgcaaaattagaagtcaaaataaacagctcccatgggcagtggtgatctcaccctggcct
ttccttttcagtgggctcagaccctcccaccgcctgctgcttttcttacaccgcgaggaagcttcct
cgcaactttgtggtagattactatgagaccagcagcctctgctcccagccagctgtgggtgtgagta
tcaacccttggtgacctgggaggcaagggtgagggctggatttttaaagggggacctgttttgggg
agggggtgatgagcgtggggaggcagctctcagggctgaagccttccctgacagcagtgagggtca
cagggtcatgaactcacttttcaagtgtgaaggcggctgagtggcagccgagacagaagggggttc
ctggggagggaagttattcagaggacaggggaagcaggggaaggcagacaggtcccatgagatatgga
ccaattccttaaaccatgctagaaaaacatgtggaaaagtcactaccaggctggcagggaatgggg
caatctattcatactgattgcaatgccactgggttcctaattctgggcaaccctggggccacagc
taaattccagtgagtggaagttacagggagtcctgcttccagtgctgctcgaggaaggatcccatcca
ccagagctgccccacatggaccatgggtcaggcagaggaagatgcctaccacaggcaagggaataaag
ccagatgacctcaaagggtcccatgggattctaatctgtctgctccttgttctacagattccaaacc
aaaagaggcaagcaagtctgcgctgaccccagtgagtcctgggtccaggagtagctgtatgacctg
gaactgaactgagctgctcagagacaggaagtcttc

Example of a Human Gene

cctcacctctgagaaaacctctttgccaccaataccatgaagctctgcgtgactgtcctgtctctc
ctcgtgctagtagctgccttctgctctctagcactctcagcaccaagtaagtctacttttgcagct
gctatttctgagtcaaggtgtaggcagagtccttttttctagtcattggctggcaaacagtgggatct
ggggatgggacaaaaggcagctaggaagattgccatgtagtctgctgctaaatgtagagtctagta
gatattcagtaacattcaagttcctattttcttaagaattagcaaccagcagaggaaaacgatggg
ctggaagtcagactgttgaattggctctgcctttaattatttgttcaagcaagcccctgtccctct
ctgtgccttggtttcccatctgtcataatgaaggagtgcatgtgttctgagactgaatccagtt
ccaatcttctagattttctttctcgttcttctctgaagatccactattcagaataagactcctgtct
atgttaggtgggaatggatacaagggaccataatttgggggttctggtagctccacagggatgctcaa
tgaagatgcaaaattagaagtcaaaataaacagctcccatgggcagtggtgatctcaccctggcct
ttccttttcagtgggctcagaccctcccaccgctgctgctttttcttacaccgcgaggaagcttcct
cgcaactttgtggtagattactatgagaccagcagcctctgctcccagccagctgtggtgtgagta
tcaacccttggtgcccctgggaggcaagggtgagggctggatttttaaagggggcctgttttgggg
agggggtgatgagcgtggggaggcagctctcagggtgaagccttccctgacagcagtgagggtca
cagggtcatgaactcacttttcaagtgtgaaggcggctgagtggcagccgagacagaagggggttc
ctggggaggaagttattcagaggacagggaagcaggggaaggcagacagggtcccatgagatatgga
ccaattccttaaaccatgctagaaaaacatgtggaaaagtcactaccagggtggcagggaatgggg
caatctattcatactgattgcaatgccactgggttcctaatctgggcaaccctggggccacagc
taaattccagtgagtggaaagtacaggagagtcctgcttccagtgctgctcgaggaaggatcccatcca
ccagagctgccccacatggaccatgggtcaggcagaggaagatgcctaccacaggcaagggaataaag
ccagatgacctcaaagggtcccatgggattctaatctgtctgctccttgttctacagattccaaacc
aaaagaggcaagcaagtctgcgctgacccagtgagtcctgggtccaggagtacgtgtatgacctg
gaactgaactgagctgctcagagacaggaagtcttc

Methods of Gene Prediction



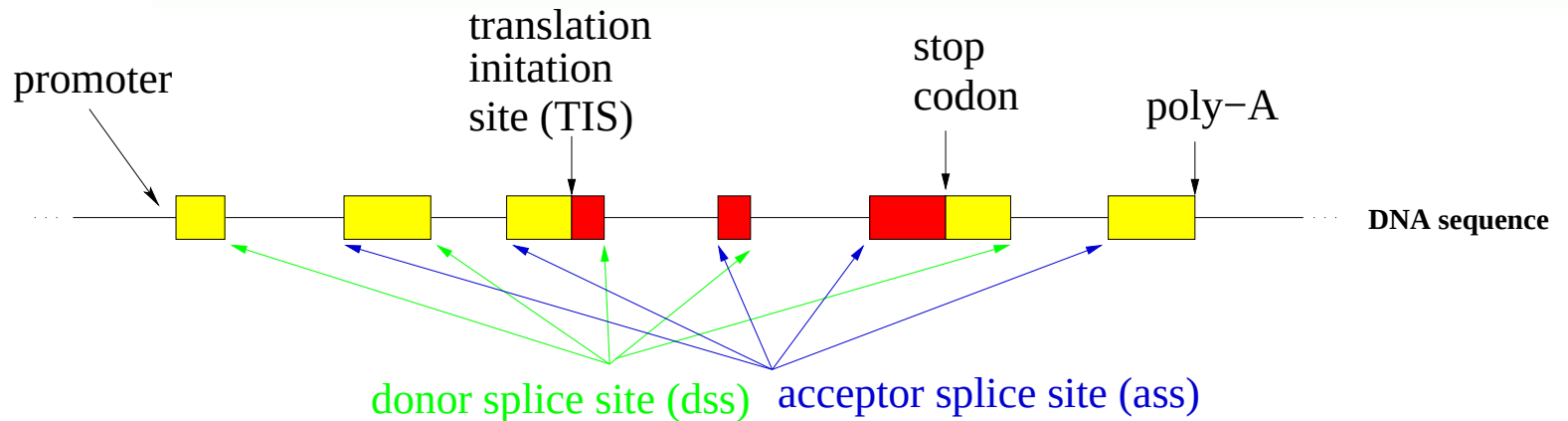
- ⑥ signal sensors
- ⑥ content sensors
- ⑥ alignment with cDNA
- ⑥ alignment with ESTs
- ⑥ protein homology
- ⑥ cross-species DNA comparison

Methods of Gene Prediction



- ⑥ signal sensors
 - ⑥ content sensors
 - ⑥ alignment with cDNA
 - ⑥ alignment with ESTs
 - ⑥ protein homology
 - ⑥ cross-species DNA comparison
- } ab initio methods
- } extrinsic methods

Signal Sensors



Signals are short sequence segments of the DNA, that control translation or transcription.

Signal Sensors

- ⑥ promoter: marks the begin of transcription
- ⑥ splice sites: 5' (donor) and 3' (acceptor) end of an intron
- ⑥ TIS: Contains the start codon (usually atg) and marks the begin of translation
- ⑥ stop codon: marks the end of translation (usually tga, taa or tag)
- ⑥ poly-A signal: triggers end of transcription.

Donor Splice Site



These signals contain **typical** sequence motifs, but these motifs are not **characteristic**: The motifs occur also at positions where actually no signal is.

Example donor splice sites: (Almost) every intron begins with the dinucleotide **gt**, but that is not sufficiently specific and does not suffice for locating donor splice sites.

Donor Splice Site

cctcacctctgagaaaacctctttgccaccaataccatgaagctctgctgtgactgtcctgtctctc
ctcgtgctagtagctgccttctgctctcttagcactctcagcaccaagttaaagtctacttttgcagct
gctatttctgagtcaagggtgtaggcagagtccttttttctagtcattggctggcaaacagtgggatct
ggggatgggacaaaaggcagctaggaagattgccatgtagtctgctgctaaatgtagagtctagta
gatattcagtaaacattcaaggttcctattttcttaagaattagcaaccagcagaggaaaacgatggg
ctggaaagtcagactgttgaattggctctgctttaaattatttgttcaagcaagcccctgtccctct
ctgtgccttggtttcccatctgtcatatgaagggaagtcgatgtgttctgagactgaatccaagt
ccaatcttctagattttctttctcgttcttctctgaagatccactattcagaataagactcctgctc
atgttaggtgggaatggatacaagggaccataatttggggttctggtagctccacagggatgctcaa
tgaagatgcaaaattagaaagtcaaaataaacagctcccatgggcaagtggtgatctcacccctggcct
ttccttttcagtgggctcagaccctcccaccgcctgctgcttttcttacaccgcgaggaagcttcct
cgcaactttgtggtagattactatgagaccagcagcctctgctcccagccagctgtgggtgtgagta
tcaacccttggtgcccctgggaggcaagggtgagggctggatttttaaagggggcctgttttgggg
aggggtgatgagcgtggggaggcagctctcagggctgaagccttccctgacagcagtgaggtca
caggtcatgaactcacttttcaaagtgctgaaggcggctgagtggtgagccgagacagaagggggttc
ctggggaggaaagtattcagaggacagggaagcaggggaaggcagacaggtcccatgagatatgga
ccaattccttaaaccatgctagaaaaacatgtggaaaagtcactaccaggctggcagggaatgggg
caatctattcatactgattgcaatgccactgttcctaattctgggcaaccctggggccacagc
taaattccaagtgaagtggaagtgtacagggaagtcgttccaagtgctgctcgaggaaggatcccatcca
ccagagctgccccacatggaccatgtcaggcagaggaagatgcctaccacaggcaagggaataaag
ccagatgacctcaaaggtcccatgggattctaatctgtctgctccttgttctacagattccaaacc
aaaagaggcaagcaaagtcctgctgaccccaagtgaggtcctgggtccaggagtaggtatgacctg
gaactgaactgagctgctcagagacaggaaagtcct

Donor Splice Site

Exon

Intron

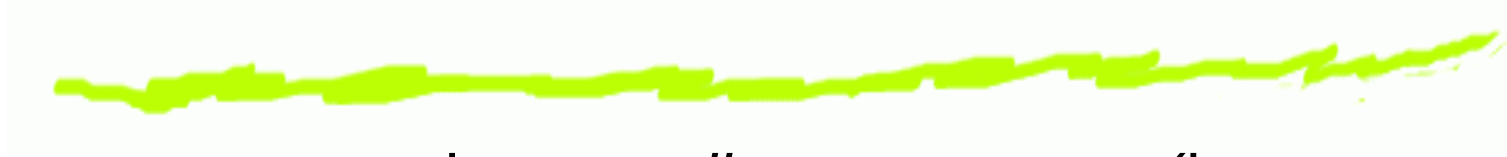
```
...agagcaaggtacgc...
...tctttcatgtgagt...
...actgtcaggtatgt...
...gacaaaaggtacgt...
...tccaaaaggcaggg...
...tttcctaggtaacg...
...agaagatgttagga...
...tcctttgggtgagt...
...gatcctgggtcagt...
...cacgctgggtacgc...
...atcacctcgtgagt...
...cttccagggtgaga...
...acctcacggtgaga...
...aacaggaggtacca...
...gtcggaccgtgagt...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

Donor Splice Site



A pictogram of the region -8 to 6 relative to the boundary between exon and intron. The size of the letters of the bases is proportional to the relative frequency of the base at this position.

Content Sensors



Coding sequences and *non-coding* sequences (introns, intergenic region) also typically have different **base compositions**. For example

- ⑥ coding: bases g and c slightly more common
- ⑥ non-coding: bases a and t slightly more common

Content Sensors

cctcacctctgagaaaacctctttgccaccaataccatgaagctctgcgtgactgtcctgtctctc
ctcgtgctagtagctgccttctgctctctagcactctcagcaccaagtaagtctacttttgcagct
gctatttctgagtcaaggtgtaggcagagtccttttttctagtcattggctggcaaacagtgggatct
ggggatgggacaaaaggcagctaggaagattgccatgtagctctgctgctaaatgtagagcttagta
gatattcagtaacattcaagttcctattttcttaagaattagcaaccagcagaggaaaacgatggg
ctggaagtcagactgttgaattggctctgcctttaattatttgttcaagcaagcccctgtccctct
ctgtgccttggtttcccatctgtcataatgaaggagtgcatgtgttctgagactgaatccagtt
ccaatcttctagattttctttctcgttcttctctgaagatccactattcagaataagactcctgtct
atgttaggtgggaatggatacaagggaccataatttgggggttctggtagctccacagggatgctcaa
tgaagatgcaaaattagaagtcaaaataaacagctcccatgggcagtggtgatctcaccctggcct
ttccttttcagtgggctcagaccctcccaccgctgctgctttttcttacaccgcgaggaagcttcct
cgcaactttgtggtagattactatgagaccagcagcctctgctcccagccagctgtggtgtgagta
tcaacccttggtgcccctgggaggcaagggtgagggtggtatttttaaagggggcctgttttgggg
agggggtgatgagcgtggggaggcagctctcagggtgaagccttccctgacagcagtgagggtca
cagggtcatgaactcacttttcaagtgtgaaggcggctgagtggcagccgagacagaagggggttc
ctggggaggaagttattcagaggacagggaagcaggggaaggcagacagggtcccatgagatatgga
ccaattccttaaaccatgctagaaaaacatgtggaaggatcactaccagggtggcagggaatgggg
caatctattcatactgattgcaatgccactgggttcctaatctgggcaaccctggggcccacagc
taaattccagtgagtggaagttacagggtgctgcttccagtgctgctcagaggaaggatcccatcca
ccagagctgccccacatggaccatggtcaggcagaggaagatgcctaccacaggcaagggaataaag
ccagatgacctcaaagggtcccatgggattctaatctgtctgctccttgttctacagattccaaacc
aaaagaggcaagcaagtctgcgctgacccagtgagtcctgggtccaggagtacgtgtatgacctg
gaactgaactgagctgctcagagacaggaagtcttc

Content Sensors



The frequency of **dinucleotides** serve even better as a means to distinguish coding from non-coding sequences. Example: The dinucleotide **at** occurs more often in non-coding sequences than in coding sequences.

Content Sensors

cctcacctctgagaaaacctctttgccaccaataccatgaagctctgcgtgactgtcctgtctctc
ctcgtgctagtagctgccttctgctctctagcactctcagcaccaagtaagtctacttttgcagct
gctatttcgagtcaaggtgtaggcagagtccttttttctagtcattggctggcaaacagtgggact
ggggatgggacaaaaggcagctaggaagattgccatgtagctctgctgctaatagttagagcttagta
gatatcagtaaacattcaagttcctattttcttaagaattagcaaccagcagaggaaaacgattggg
ctggaagtcagactgttgaattggctctgcctttaattattttgttcaagcaagcccctgtccctct
ctgtgccttggtttcccccattctgtcatatgaaggagtgcgattgtgttctgagactgaattccagtt
ccaattcttctagattttctttctcgttcttctctgaagattccactattcagaataagactcctgctc
atgttaggtgggaattggatacaagggaccatatttgggggttctggtagctccacagggaattgctca
tgaagatgcaaaaattagaagtcaaaaataaacagctcccattgggcagtgttgattctcaccctggcct
ttccttttcagtgggctcagaccctcccaccgcctgtgctttttcttacaccgcgaggaagcttcct
cgcaactttgtggtagattactatgagaccagcagcctctgctcccagccagctgtggtgtgagta
tcaacccttggtgcccctgggaggcaagggtgagggctggatttttaaagggggcctgttttgggg
aggggggtgattgagcgtgaggagggcagctctcagggctgaagccttccctgacagcagtgaggtca
caggtcattgaactcacttttcaagtgtgaaggcggctgagtggcagccgagacagaagggggttc
ctggggaggaagttattcagaggacaggggaagcaggggaaggcagacaggtcccattgagatatgga
ccaatttccttaaaccattgctagaaaaaacattgtgaaaagtcactaccaggctggcagggaattggg
caattctattcattactgattgcaattgccactgggttcctaattctgggcaaccctggggccacagc
taaattccagtgagtggaagttacagggagtcctgcttccagtgctgctcgaggaaggattcccattcca
ccagagctgccccacattggaccattggtcaggcagaggaagattgcctaccacaggcaagggaataaag
ccagattgacctcaaagggtcccattgggattctaatctgtctgctccttggttctacagattccaaacc
aaaagaggcaagcaagctctgcgctgaccccagtgagtcctgggtccaggagtacgtgtatgacctg
gaactgaactgagctgctcagagacaggaagtcttc

Content Sensors

Reading frame dependent **hexamer** frequencies is the most commonly used **content sensor** of current gene prediction programs.

Protein Homology

Use **local similarity** between translated input DNA sequence and amino acid sequence from database to **infer evidence** about coding regions.

Example:

input human DNA:	gcc atg tcg tcc ggc atc cat gta gcg ctg gtg act gga ggc aac aag ggc atc ggc ...
translated 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 ...

Programs:

- ⑥ GenomeScan
- ⑥ AUGUSTUS+

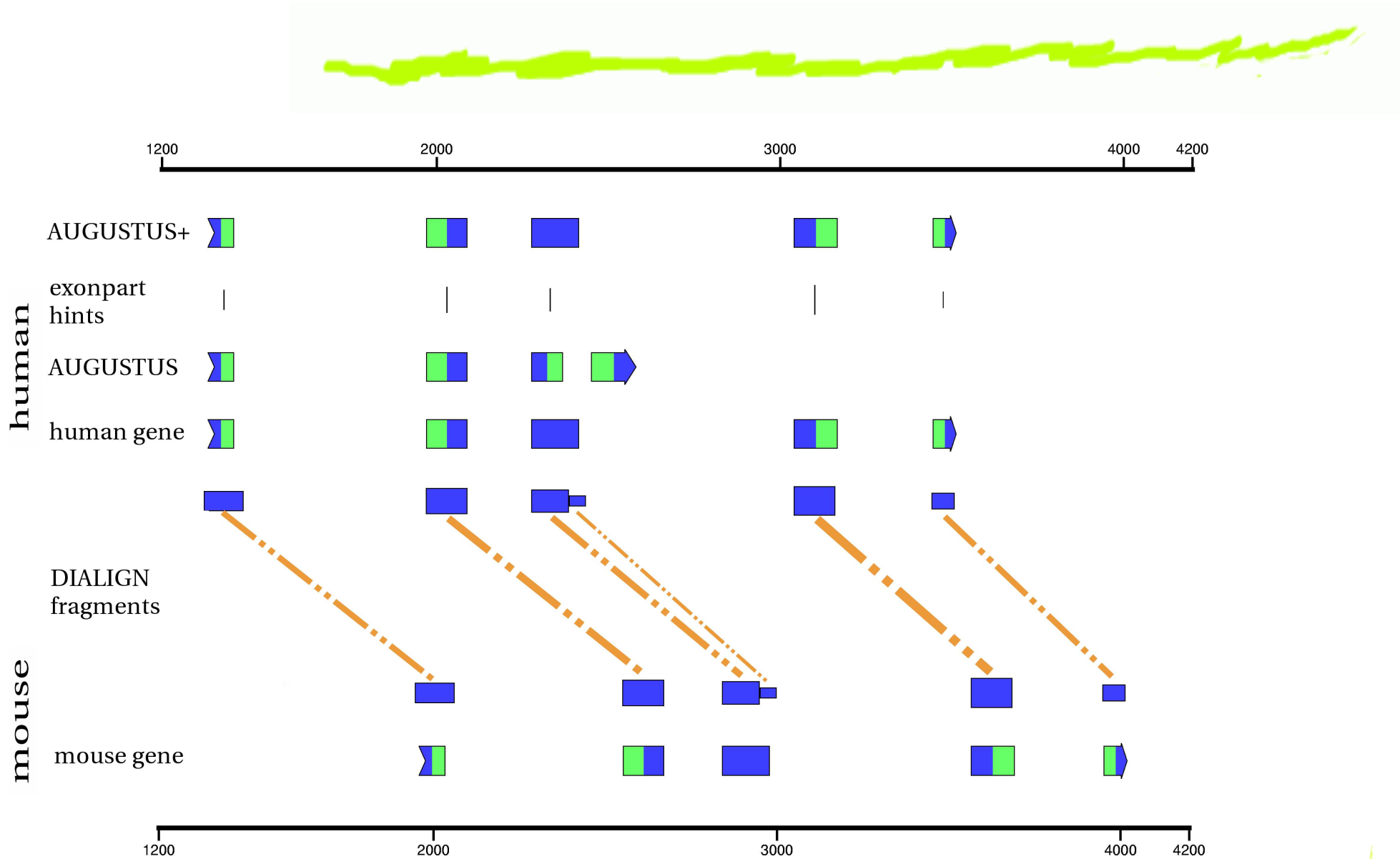
Cross-Species DNA Comparison

Consider the DNA sequences of two different species coding for the 'same' (or a similar) protein. Functional parts of the sequence, especially coding regions, tend to be more conserved.

Programs:

- ⑥ DoubleScan
- ⑥ SLAM
- ⑥ AUGUSTUS+
- ⑥ AGenDA

Cross-Species DNA Comparison



The Reliability of Gene Prediction ...

... depends on the organism and on the available extrinsic information.

Example: humans, ab initio, ENCODE sequences

program	base		exon		transcr.		gene	
	sn	sp	sn	sp	sn	sp	sn	sp
Augustus	78.7	75.3	52.4	62.9	11.1	17.2	24.3	17.2
GenemarkHMM	78.4	38.0	50.6	29.0	6.9	3.2	15.2	3.2
Genezilla	87.6	50.9	62.1	50.3	9.1	8.8	19.6	8.8
geneid	76.8	76.5	53.8	61.1	4.8	8.8	10.5	8.8
genscan	84.2	60.6	58.7	46.4	7.4	10.1	15.5	10.1

Problems of Gene Prediction

- ⑥ alternative splicing
- ⑥ TIS and promoter prediction less reliable
- ⑥ exceptions, e.g.
 - △ overlapping genes
 - △ non-canonical splice sites
 - △ (programmed) frameshifts
 - △ different genetic code

Problems of Gene Prediction

- ⑥ alternative splicing
- ⑥ TIS and promoter prediction less reliable
- ⑥ exceptions, e.g.
 - △ overlapping genes
 - △ non-canonical splice sites
 - △ (programmed) frameshifts
 - △ different genetic code

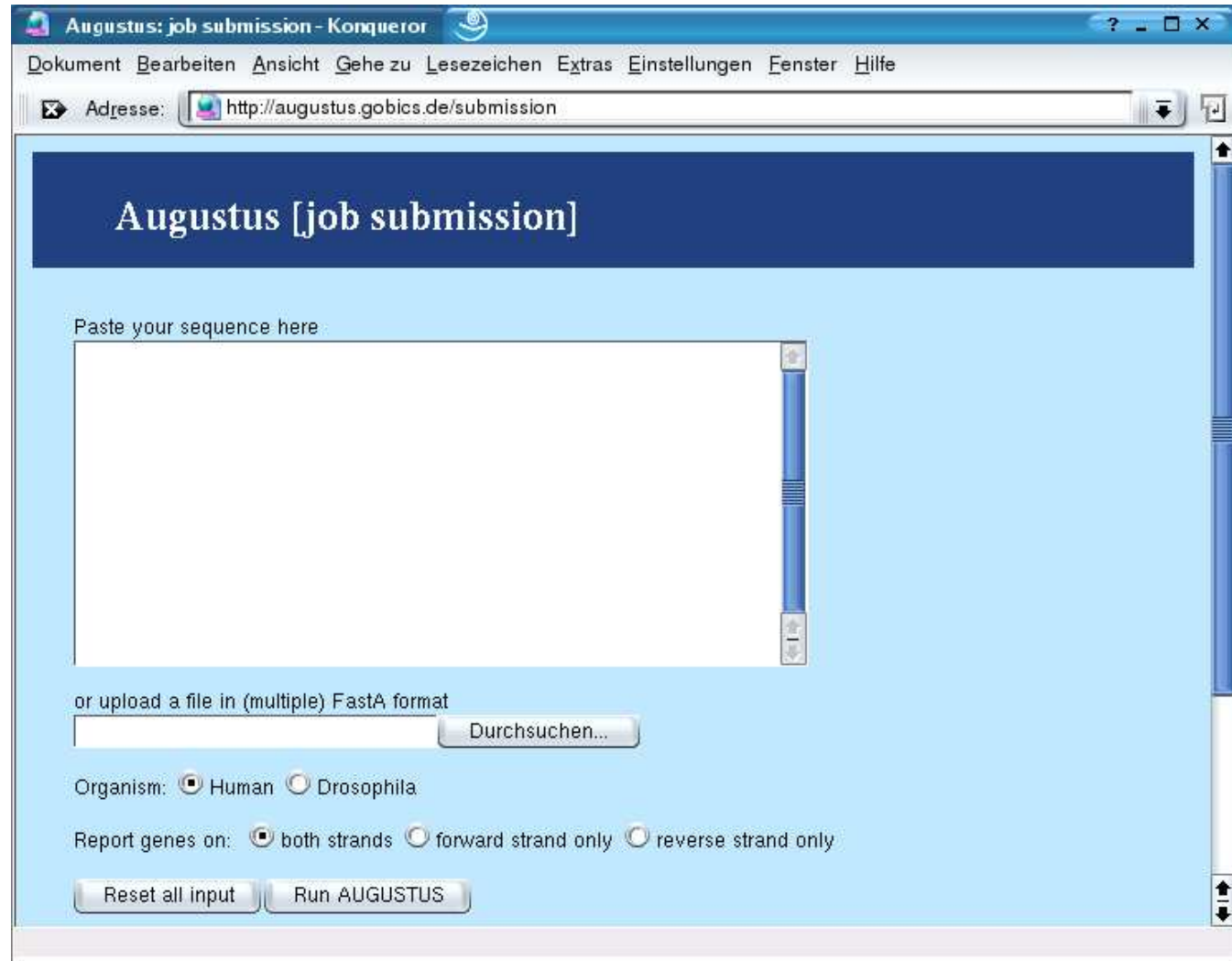
Problems of Gene Prediction

- ⑥ alternative splicing
- ⑥ TIS and promoter prediction less reliable
- ⑥ exceptions, e.g.
 - △ overlapping genes
 - △ non-canonical splice sites
 - △ (programmed) frameshifts
 - △ different genetic code

Problems of Gene Prediction

- ⑥ alternative splicing
- ⑥ TIS and promoter prediction less reliable
- ⑥ exceptions, e.g.
 - △ overlapping genes
 - △ non-canonical splice sites
 - △ (programmed) frameshifts
 - △ different genetic code

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:". To the right of this field is a vertical scrollbar. Below the text field, there is a label "or upload a file in (multiple) FastA format" followed by a file selection button labeled "Durchsuchen...". Below the file selection button, there are two radio buttons for "Organism": "Human" (selected) and "Drosophila". Below the organism selection, there are three radio buttons for "Report genes on": "both strands" (selected), "forward strand only", and "reverse strand only". At the bottom of the form, 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