Genannotation bei Prokaryoten

Maike Tech

Abt. Bioinformatik
Institut für Mikrobiologie und Genetik (IMG)
Universität Göttingen

28. November 2005

Genetik von Pro- und Eukaryoten

Eukaryoten	**Prokaryoten**
Zellkern	kein Zellkern
Intron-/Exonstruktur	Single-Exon
–	Operon-Strukturen
ca. 10% kodierend	ca. 80% kodierend
mehrere Mrd. Basen	mehrere Mio. Basen
einige 10-100tausend Gene	einige tausend Gene

Genexpression bei Prokaryoten

DNA

⇓ Transkription (RNA-Pol)

mRNA

⇓Translation (Ribosom)

AS-Seq.

- Promotor
- Ribosombindestelle mit SD-Sequenz (Shine-Dalgarno-Sequenz)
- Translationsstart (z.B. AUG bzw. Met)
- kodierende Region (Gen)
- Translationsstop (z.B. UGA)
- Terminator

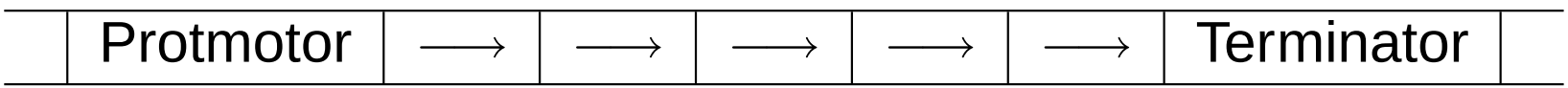
Kodierung durch Tripletts

		2					
		T/U	C	A	G		
1	T/U	TTT (Phe)	TCT (Ser)	TAT (Tyr)	TGT (Cys)	T/U	3
		TTC (Phe)	TCC (Ser)	TAC (Tyr)	TGC (Cys)	C	
		TTA (Leu)	TCA (Ser)	TAA (Stop)	TGA (Stop)	A	
		TTG (Leu)	TCG (Ser)	TAG (Stop)	TGG (Trp)	G	
	C	CTT (Leu)	CCT (Pro)	CAT (His)	CGT (Arg)	T/U	
		CTC (Leu)	CCC (Pro)	CAC (His)	CGC (Arg)	C	
		CTA (Leu)	CCA (Pro)	CAA (Gln)	CGA (Arg)	A	
		CTG (Leu)	CCG (Pro)	CAG (Gln)	CGG (Arg)	G	
	A	ATT (Ile)	ACT (Thr)	AAT (Asn)	AGT (Ser)	T/U	
		ATC (Ile)	ACA (Thr)	AAC (Asn)	AGC (Ser)	C	
		ATA (Ile)	ACA (Thr)	AAA (Lys)	AGA (Arg)	A	
		ATG (Met)	ACG (Thr)	AAG (Lys)	AGG (Arg)	G	
	G	GTT (Val)	GCT (Ala)	GAT (Asp)	GGT (Gly)	T/U	
		GTC (Val)	GCC (Ala)	GAC (Asp)	GGC (Gly)	C	
		GTA (Val)	GCA (Ala)	GAA (Glu)	GGA (Gly)	A	
		GTG (Val)	GCG (Ala)	GAG (Glu)	GGG (Gly)	G	

Operon-Struktur

Alle Gene des Operons werden gemeinsam transkribiert (resultieren also in einer mRNA). Die Trennung der Produkte erfolgt erst durch die Ribosomen. Daher sind alle Gene eines Operons gleich orientiert und die Abstände zwischen den einzelnen sind sehr gering (oft 1,4 oder 8 BP).

Aufbau:



Von der DNA zur Annotation

1. **Sequenzierung:** Mit Shotgun-Methode mechanisches Zerkleinern der Sequenz, dann Replikation der Fragmente mit Kettenabbruch-Methode, Resultat sind Fragmente unterschiedlicher Länge
2. **Basecalling:** Die replizierten Sequenzstücke werden im elektrischen Feld getrennt. Die Enden sind mit Fluoreszensfarbstoffen gelabelt, dadurch kann jeweils die letzte Base identifiziert werden.
3. **Assemblierung:** Zusammenfügen der Fragmente
4. **Annotation:** Genvorhersage, Zuordnung einer Funktion

Von der DNA zur Annotation

1. **Sequenzierung:** Mit Shotgun-Methode mechanisches Zerkleinern der Sequenz, dann Replikation der Fragmente mit Kettenabbruch-Methode, Resultat sind Fragmente unterschiedlicher Länge
2. **Basecalling:** Die replizierten Sequenzstücke werden im elektrischen Feld getrennt. Die Enden sind mit Fluoreszensfarbstoffen gelabelt, dadurch kann jeweils die letzte Base identifiziert werden.
3. **Assemblierung:** Zusammenfügen der Fragmente
4. **Annotation:** Genvorhersage, Zuordnung einer Funktion

Von der DNA zur Annotation

1. **Sequenzierung:** Mit Shotgun-Methode mechanisches Zerkleinern der Sequenz, dann Replikation der Fragmente mit Kettenabbruch-Methode, Resultat sind Fragmente unterschiedlicher Länge
2. **Basecalling:** Die replizierten Sequenzstücke werden im elektrischen Feld getrennt. Die Enden sind mit Fluoreszensfarbstoffen gelabelt, dadurch kann jeweils die letzte Base identifiziert werden.
3. **Assemblierung:** Zusammenfügen der Fragmente
4. **Annotation:** Genvorhersage, Zuordnung einer Funktion

Von der DNA zur Annotation

1. **Sequenzierung:** Mit Shotgun-Methode mechanisches Zerkleinern der Sequenz, dann Replikation der Fragmente mit Kettenabbruch-Methode, Resultat sind Fragmente unterschiedlicher Länge
2. **Basecalling:** Die replizierten Sequenzstücke werden im elektrischen Feld getrennt. Die Enden sind mit Fluoreszensfarbstoffen gelabelt, dadurch kann jeweils die letzte Base identifiziert werden.
3. **Assemblierung:** Zusammenfügen der Fragmente
4. **Annotation:** Genvorhersage, Zuordnung einer Funktion

Nucleotidsequenz im FASTA-Format

6 Leserahmen, 3 im »+«-Strang, 3 im »–«-Strang

```
>gi|6626251|Escherichia coli K-12 complete genome
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAG
AGTGTCTGATAGCAGCTTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAATTT
TATTGACTTAGGTCACTAAATACTTTAACCAATATAGGCATAGCGCACAGACAG
ATAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACCATTACC
ACCACCATCACCATTACCACAGGTAACGGTGCGGGCTGACGCGTACAGGAAACA
CAGAAAAAAGCCCGCACCTGACAGTGCGGGCTTTTTTTTTCGACCAA ...
```

Nucleotidsequenz im FASTA-Format

6 Leserahmen, 3 im »+«-Strang, 3 im »–«-Strang

```
>gi|6626251|Escherichia coli K-12 complete genome
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAG
AGTGTCTGATAGCAGCTTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAATTT
TATTGACTTAGGTCACTAAATACTTTAACCAATATAGGCATAGCGCACAGACAG
ATAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACCATTACC
ACCACCATCACCATTACCACAGGTAACGGTGCGGGCTGACGCGTACAGGAAACA
CAGAAAAAAGCCCGCACCTGACAGTGCGGGCTTTTTTTTTCGACCAA ...
```

Nucleotidsequenz im FASTA-Format

6 Leserahmen, 3 im »+«-Strang, 3 im »–«-Strang

```
>gi|6626251|Escherichia coli K-12 complete genome
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAG
AGTGTCTGATAGCAGCTTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAATTT
TATTGACTTAGGTCACTAAATACTTTAACCAATATAGGCATAGCGCACAGACAG
ATAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACCATTACC
ACCACCATCACCATTACCACAGGTAACGGTGCGGGCTGACGCGTACAGGAAACA
CAGAAAAAAGCCCGCACCTGACAGTGCGGGCTTTTTTTTTCGACCAA ...
```

Nucleotidsequenz im FASTA-Format

6 Leserahmen, 3 im »+«-Strang, 3 im »–«-Strang

```
>gi|6626251|Escherichia coli K-12 complete genome
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAG
AGTGTCTGATAGCAGCTTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAATTT
TATTGACTTAGGTCACTAAATACTTTAACCAATATAGGCATAGCGCACAGACAG
ATAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACCATTACC
ACCACCATCACCATTACCACAGGTAACGGTGCGGGCTGACGCGTACAGGAAACA
CAGAAAAAAGCCCGCACCTGACAGTGCGGGCTTTTTTTTTCGACCAA ...
```

Vorgehensweise

- Suche nach »sicheren« Kandidaten ORFs, Erstellen eines Modells
- Vorhersage wahrscheinlich kodierender Regionen. Dabei werden in den meisten prokaryotischen Genomen 98%–99% der kodierenden Regionen gefunden.
- Annotation des wahrscheinlichsten Starts
- Entfernen von Überlappungen

Definition eines ORFs

Ein ORF (**O**pen **R**eading **F**rame) ist definiert als Anzahl n unmittelbar aufeinanderfolgender Tripletts $t_1 ... t_n$, wobei t_1 das Startcodon und t_n das Stopcodon ist. Jeder ORF beginnt mit einem Startcodon aus der Menge {ATG, GTG, TTG, CTG} und endet mit dem nächsten terminalen Codon aus der Menge {TAG, TAA, TGA}, das im gleichen Leserahmen liegt. Wenn ein ORF kodiert, wird er als Gen bezeichnet.

Vorgehensweise

- Suche nach »sicheren« Kandidaten ORFs, Erstellen eines Modells
- Vorhersage wahrscheinlich kodierender Regionen. Dabei werden in den meisten prokaryotischen Genomen 98%–99% der kodierenden Regionen gefunden.
- Annotation des wahrscheinlichsten Starts
- Entfernen von Überlappungen

Vorgehensweise

- Suche nach »sicheren« Kandidaten ORFs, Erstellen eines Modells
- Vorhersage wahrscheinlich kodierender Regionen. Dabei werden in den meisten prokaryotischen Genomen 98%–99% der kodierenden Regionen gefunden.
- Annotation des wahrscheinlichsten Starts
- Entfernen von Überlappungen

Vorgehensweise

- Suche nach »sicheren« Kandidaten ORFs, Erstellen eines Modells
- Vorhersage wahrscheinlich kodierender Regionen. Dabei werden in den meisten prokaryotischen Genomen 98%–99% der kodierenden Regionen gefunden.
- Annotation des wahrscheinlichsten Starts
- Entfernen von Überlappungen

Probleme bei der Vorhersage

- Annotation der Translationsstarts, da *Startcodons* auch innerhalb von Genen kodieren können
- Hohe Rate »Falsch Positiver«, d. h. menschliche Annotatoren müssen die Vorhersage überprüfen
- Frame-Shifts können auftreten, d. h. Start und Stop liegen nicht im selben Leserahmen (selten)
- Schlechte Vorhersagen für »heterogene« Genome

Vorhersage kodierender Regionen

1. extrinsisch
   - Suche nach Ähnlichkeiten zu bereits bekannten Genen in Datenbanken meist mit BLAST (**B**asic **L**ocal **A**lignment **S**earch **T**ool) oder ähnlichen Programmen.
2. intrinsisch (*ab initio*)
   - Erstellen von Modellen mit Hilfe der »initialen« ORFs für kodierende-/nicht kodierende Regionen, z. B. basierend auf Häufigkeiten von *k*meren

BLAST-Suche

```
>thrA_337_2799_+
          Length = 2463
 Score =  509 bits (257), Expect = e-145
 Identities = 599/713 (84%)
 Strand = Plus / Plus

 Query: 337  atgcgagtgttgaagttcggcggtacatcagtggcaaatgcagaacgttttctgcgtgtt 396
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
 Sbjct: 1    atgcgagtgttgaagttcggcggtacatcagtggcaaatgcagaacgttttctgcgtgtt 60


 Query: 397  gccgatattctggaaagcaatgccaggcaagggcaggtagcgaccgtactttccgccccc 456
              ||||||||||||||||||||||||| |||||||| || ||||| || || ||||||
 Sbjct: 61   gccgatattctggaaagcaatgccaggcaggggcaggtggccaccgtcctctctgccccc 120

 ...
```

Vorhersage kodierender Regionen

1. extrinsisch
   - Suche nach Ähnlichkeiten zu bereits bekannten Genen in Datenbanken meist mit BLAST (**B**asic **L**ocal **A**lignment **S**earch **T**ool) oder ähnlichen Programmen.
2. intrinsisch (*ab initio*)
   - Erstellen von Modellen mit Hilfe der »initialen« ORFs für kodierende-/nicht kodierende Regionen, z. B. basierend auf Häufigkeiten von *k*meren

Häufigkeiten

Es werden die Häufigkeiten von Oligomeren in kodierenden und nicht-kodierenden Regionen gezählt. Die Häufigkeiten können dabei positionsabhängig und positionsunabhängig festgestellt werden. Enthält die Sequenz, die untersucht wird, viele Oligomere, die in kodierenden Regionen auftreten, so ist es wahrscheinlich, daß diese Sequenz auch kodierend ist.
Ein Score S kann beispielsweise folgendermaßen berechnet werden:

$$S = \frac{f_{kodierend}(n\text{mer})}{f_{nicht-kodierend}(n\text{mer})}$$

Markow-Modelle

- Feststellen der Häufigkeiten von Oligomeren der Länge $k+1$ in kodierenden und nicht-kodierenden Regionen
- Darauf basiert das Modell:

$$p_{kodierend}(b_{k+1}|b_{1\ldots k}) = \text{Wahrscheinlichkeit mit der } b_{k+1} \text{ auf } b_{1..k} \text{ in kodierenden Regionen folgt}$$

$$p_{nicht-kodierend}(b_{k+1}|b_{1\ldots k}) = \text{Wahrscheinlichkeit mit der } b_{k+1} \text{ auf } b_{1\ldots k} \text{ in nicht-kodierenden Regionen folgt}$$

- für jede Sequenz S werden Wahrscheinlichkeiten $P_{kodierend}(S)$ und $P_{nicht-kodierend}(S)$ als Produkte der Einzelwahrscheinlichkeiten berechnet.

Methoden des Maschinellen Lernens

1. Die ORFs können als Vektoren in Hochdimensionalen Räumen dargestellt werden. Für die dann eine Klassifikationsfunktion gelernt wird. Mit Hilfe der Funktion können dann »ungesehene« ORFs klassifiziert werden (überwachtes Lernen).
2. Unüberwachte Verfahren (z. B. Clusterverfahren) zur Klassifikation der ORFs.

Methoden der Startvorhersage

1. DB-Suche
   - Nur für Gene mit Ähnlichkeit zu bereits bekannten, fehleranfällig (transitive Fehlerfortpflanzung)
2. Suche nach Signalen wie Ribosombindestellen (RBS – **R**ibosom **B**inding **S**ites)
   - nicht jedes Gen hat eine RBS (Operon-Organisation)
   - variiert bei verschiedenen Organismen
   - nicht für jeden Organismus bekannt (kann beispielsweise mit Hilfe der TOMPA-Methode ermittelt werden)

Signale in der mRNA

SD-Sequenz: RGGRGGTGAT (R = A oder G)

Überlappungen

Beispiel: A hat einen höheren Score als B.

Verschiebung des Startes von B	A $5' \longrightarrow 3'$ B $5' \longrightarrow 3'$
Verschiebung des Startes von A	A $5' \longrightarrow 3'$ B $5' \longrightarrow 3'$
Der Start von B wird solange verschoben, bis die Region von einen höheren Score als die von A aufweist oder die Überlappung aufgelöst ist.	$3' \longleftarrow 5'$ A B $5' \longrightarrow 3'$
B wird verworfen oder beide werden gekennzeichnet.	A $5' \longrightarrow 3'$ $3' \longleftarrow 5'$ B

Überlappungen

Beispiel: A hat einen höheren Score als B.

Verschiebung des Startes von B	A $5' \longrightarrow 3'$ B $5' \longrightarrow 3'$
Verschiebung des Startes von A	A $5' \longrightarrow 3'$ B $5' \longrightarrow 3'$
Der Start von B wird solange verschoben, bis die Region von einen höheren Score als die von A aufweist oder die Überlappung aufgelöst ist.	$3' \longleftarrow 5'$ A B $5' \longrightarrow 3'$
B wird verworfen oder beide werden gekennzeichnet.	A $5' \longrightarrow 3'$ $3' \longleftarrow 5'$ B

Überlappungen

Beispiel: A hat einen höheren Score als B.

Verschiebung des Startes von B	$A \quad 5' \longrightarrow 3'$ $B \quad 5' \longrightarrow 3'$
Verschiebung des Startes von A	$A \quad 5' \longrightarrow 3'$ $B \quad 5' \longrightarrow 3'$
Der Start von B wird solange verschoben, bis die Region von einen höheren Score als die von A aufweist oder die Überlappung aufgelöst ist.	$3' \longleftarrow 5' \quad A$ $B \quad 5' \longrightarrow 3'$
B wird verworfen oder beide werden gekennzeichnet.	$A \quad 5' \longrightarrow 3'$ $3' \longleftarrow 5' \quad B$

Überlappungen

Beispiel: A hat einen höheren Score als B.

Verschiebung des Startes von B	$A \ {}^{5'}\!\longrightarrow^{3'}$ $B \ {}^{5'}\!\longrightarrow^{3'}$
Verschiebung des Startes von A	$A \ {}^{5'}\!\longrightarrow^{3'}$ $B \ {}^{5'}\!\longrightarrow^{3'}$
Der Start von B wird solange verschoben, bis die Region von einen höheren Score als die von A aufweist oder die Überlappung aufgelöst ist.	${}^{3'}\!\longleftarrow^{5'} \ A$ $B \ {}^{5'}\!\longrightarrow^{3'}$
B wird verworfen oder beide werden gekennzeichnet.	$A \ {}^{5'}\!\longrightarrow^{3'}$ ${}^{3'}\!\longleftarrow^{5'} \ B$