

Freitag
22. April 2005
9:15

Bioinformatik Ringvorlesung Sommersemester 2005

TP3 – INF 580

t.beissbarth@dkfz.de

Raum 2.109

Tel. 42 4709

Dr. Tim Beißbarth

Deutsches Krebsforschungszentrum

Molekulare Genomanalyse

Bioinformatik

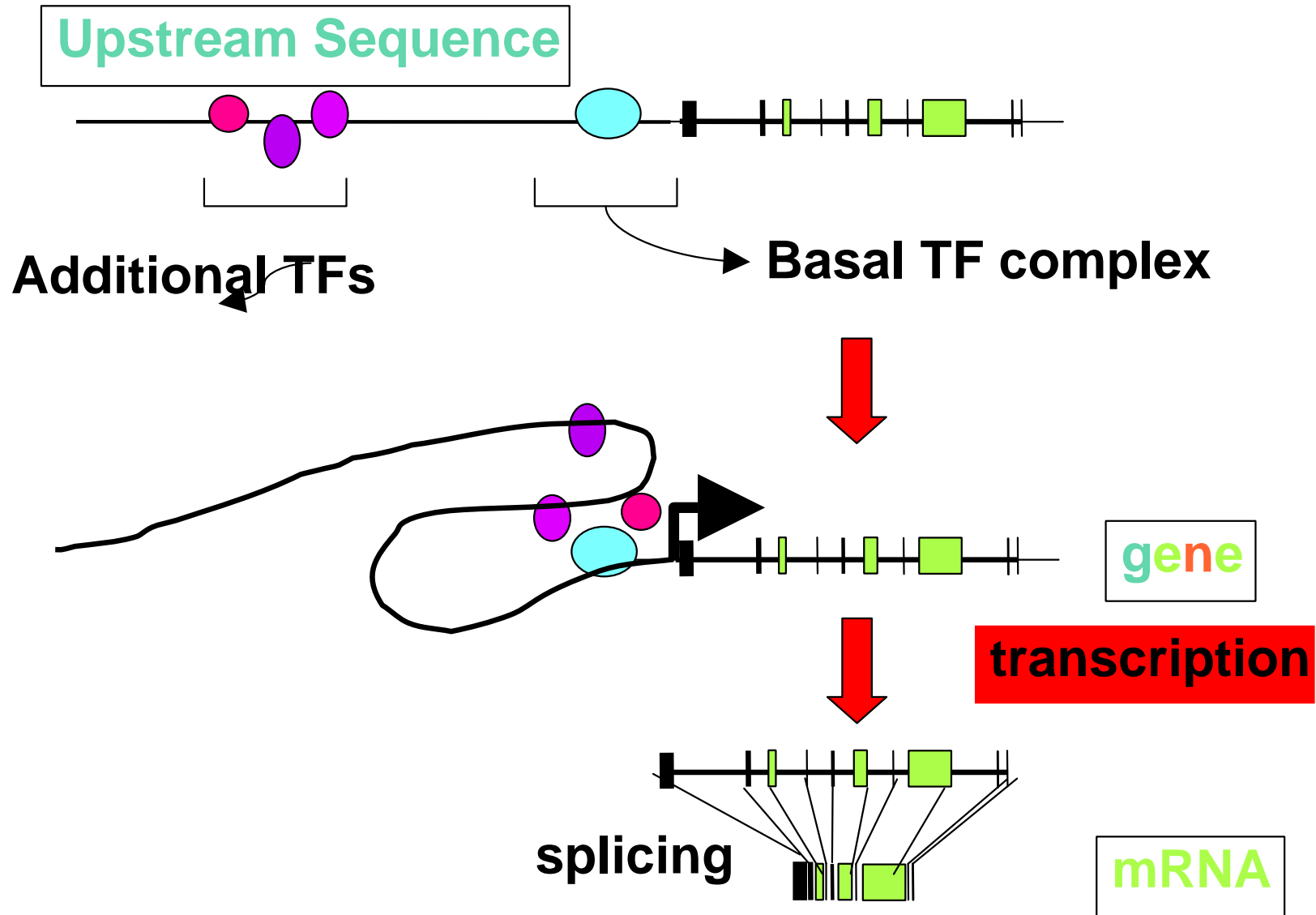
MGA

Molekulare Genomanalyse -
Bioinformatik und Datenanalyse

dkfz.

DEUTSCHES
KREBSFORSCHUNGSZENTRUM
IN DER HELMHOLTZ-GEMEINSCHAFT

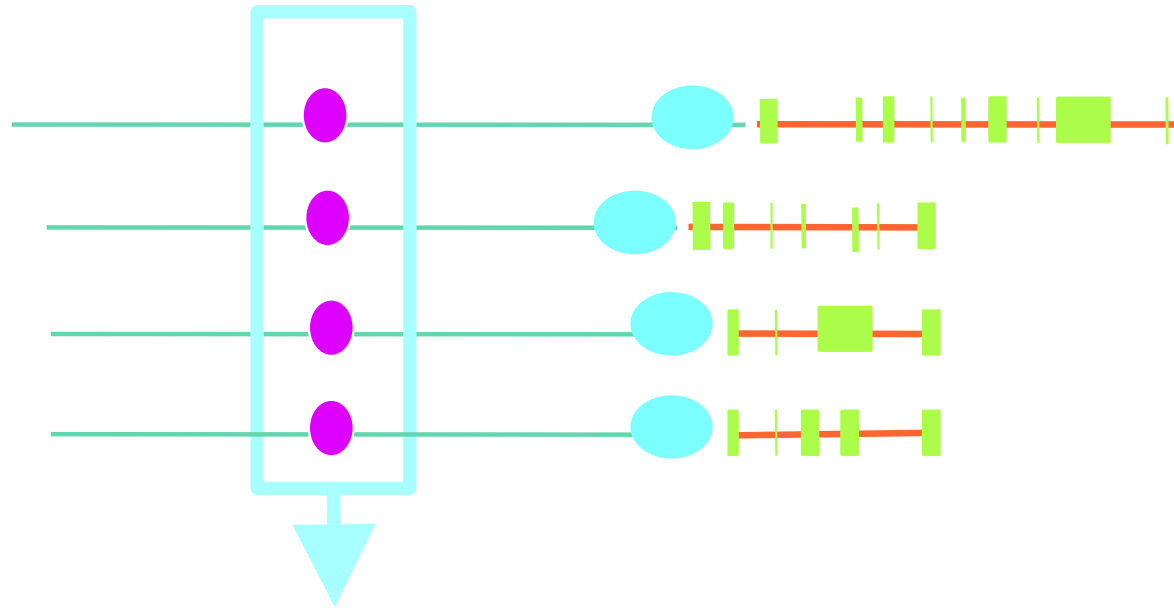
Mechanism of Transcription Initiation



Promotoren finden

- Es gibt eine Menge an Methoden um Promotoren anhand der genomischen Sequenz *in silico* vorherzusagen.
 - Suchen nach bestimmten Standardelementen – e.g. TATAAT an Position -10, TTGACA an Position -35. (*Mulligan/McClure, 1986, Nucleic Acids Research*)
 - Suchen nach gehäuftem Auftreten von Transcription Factor Binding Sites (TFBS). (*Staden, 1984, Nucleic Acids Research*)
 - Neuronale Netze (*Pedersen, 1996, ISMB*)
 - etc.
- Funktionieren alle nicht so besonders gut.

Beschreibung von Transkriptionsfaktor Bindestellen (TFBS).

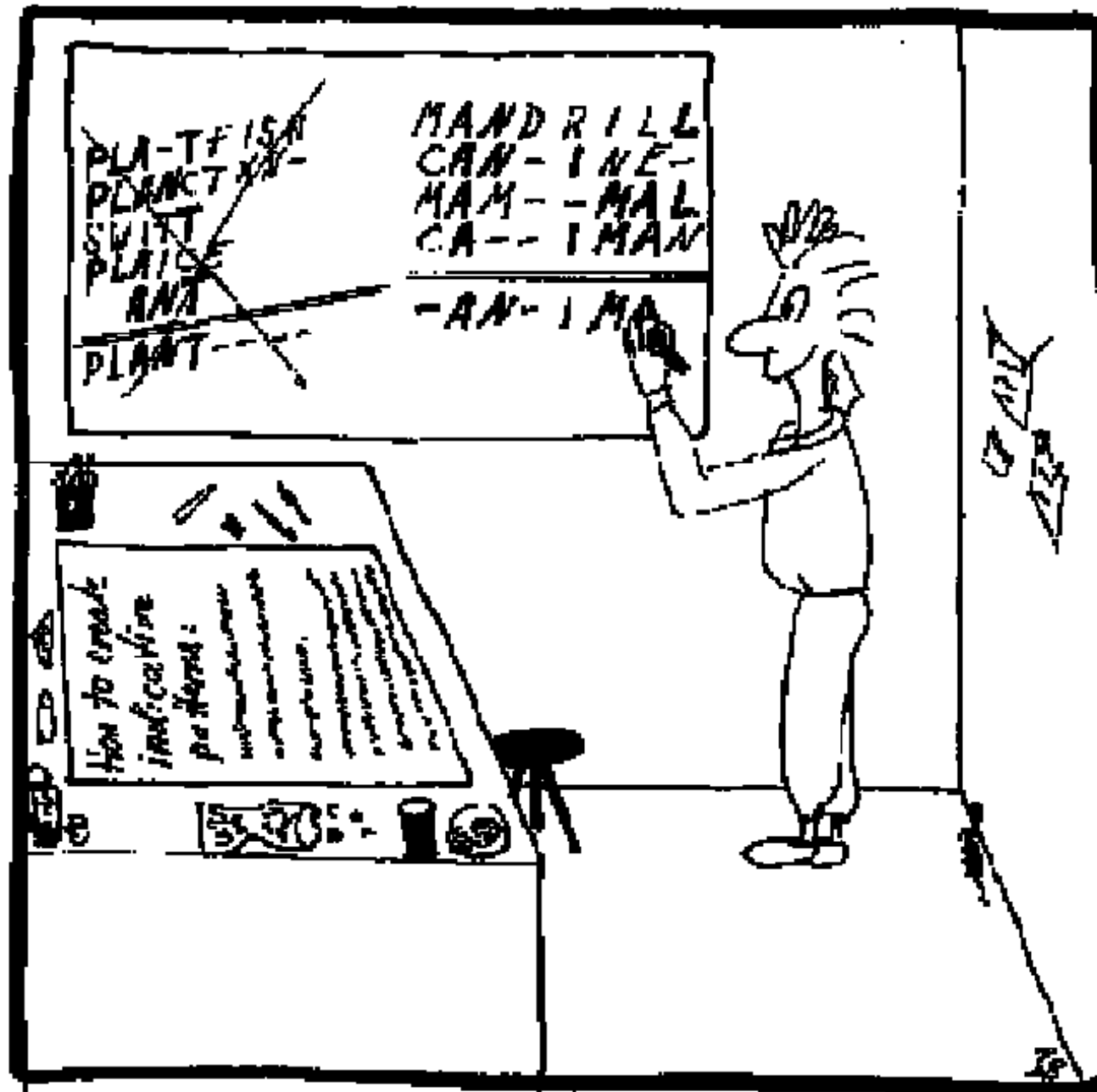


====> ATAGCTGACTA
====> TGCACTGATTA
====> TTCACTGACAG
====> AGCTCTGAATG

**Experimentell
bestimmte TFBS**

Core Region

Sequenz Motive



Brigitte Bockmann / 1995

Verschiedene Methoden zum Suchen nach Motiven (Vor-/Nachteile)

- Consensus Sequenz
Sehr einfach. Intuitiv. Einfach zu Suchen. Unflexibel. Vereinfacht.
- Regulärer Ausdruck
Einfach/Intuitiv. Einfach zu Suchen. Flexibel. Alle gefundenen Varianten gleich Wahrscheinlich. Vereinfacht.
- Position Weight Matrix (PWM, PSSM)
Statistik an einzelnen Positionen. Keine Elemente mit variabler Länge möglich.
- Hidden Markov Model
Am Flexibelsten. Kompliziert/unintuitiv. Keine Abhängigkeit zwischen Positionen möglich.
- 2nd Order Markov Model
Abhängigkeit zwischen zwei benachbarten Positionen möglich. Braucht mehr Trainingsdaten.

01. TGTGGT
 02. TGCGGT
 03. TGTGGT
 04. GGTGGT
 05. AGTGGT
 06. AGTGGT
 07. TGTGGT
 08. TATGGT
 09. TGTGGT
 10. TGTGGT
 11. TCTGGT
 12. TGCGGT
 13. TGTGGC
 14. TGCGGT
 15. TGAGGT
 16. TGTGGT
 17. TGTGGT
 18. TGTGGA
 19. TGTGGT
 20. AGTGGT
 21. TGTGGT
 22. AGTGGT
 23. TGCGGT
 24. TGTGGC
 25. TGCGGT
 26. TGTGGT
 27. TGCGGT
 28. TCTGGT
 29. TGTGGT
 30. TGTGGT
 31. TGTGGT
 32. TATGGT
 33. TGAGGT
 34. CGTGGT
 35. TGTGGT
 36. TGAGGT
 37. TGTGGT
 ...
 57. TGAGGT

Whas ist ein PSSM?

- PSSM = Position Specific Scoring Matrix = PWM = Position Weight Matrix
- Beispiel: Matrix V\$AML1_01 aus TRANSFAC

Zähle Anzahl der Nukleotide ($n_{i,j}$) an jeder Position in N alignierten Sequenzen.

Berechne Frequenz:

$$f_{i,j} = n_{i,j} / N$$

Berechne log-likelihood per pos.:

$$f_{i,j} = \ln(((n_{i,j} + p_i) / (N+1)) / p_i)$$

$$\sim \ln(f_{i,j} / p_i)$$

$p_i = a priori$ Wahrscheinlichkeit von Buchstabe 0.25

Count-Matrix							
	T	G	T	G	G	T	
A	5	2	4	0	1	1	
C	1	2	14	0	0	4	
G	2	52	1	57	55	0	
T	49	1	38	0	1	52	

Frequency-Matrix							
	T	G	T	G	G	T	
A	.09	.04	.07	.00	.02	.02	
C	.02	.04	.25	.00	.00	.07	
G	.04	.91	.02	1.0	.96	.00	
T	.86	.02	.67	.00	.02	.91	

Weigth-Matrix							
	T	G	T	G	G	T	
A	-1.0	-1.9	-1.2	-4.1	-2.5	-2.5	
C	-2.5	-1.9	0.0	-4.1	-4.1	-1.2	
G	-1.9	1.3	-2.5	1.4	1.3	-4.1	
T	1.2	-2.5	1.0	-4.1	-2.5	1.3	

PWM Statistik

- Die Wahrscheinlichkeit einer bestimmten Sequenz gegeben eine Frequency Matrix kann wie folgt ausgerechnet werden:

$$P(S) = P(s_1, \dots, s_n) = \prod_{i=1}^n P(s_i)$$

Frequency-Matrix						
	T	G	T	G	G	T
A	.09	.04	.07	.00	.02	.02
C	.02	.04	.25	.00	.00	.07
G	.04	.91	.02	1.0	.96	.00
T	.86	.02	.67	.00	.02	.91

- Beispiel CGAGGT:** $.02 \times .91 \times .07 \times 1.0 \times .96 \times .91 = 0.001$

- Als Vergleich wird die Wahrscheinlichkeit der Sequenz unter dem Null-Modell ausgerechnet.

Beispiel: alle Basen kommen gleich oft vor: $.25^6 = 0.0002$

PWM Statistik

- Jede einzelne Sequenz ist sehr unwahrscheinlich.
- Um zu vergleichen ob die Sequenz unter dem TFBS-Modell oder unter dem Null-Modell wahrscheinlicher ist berechnet man die Likelihood Ratio.

$$LR(s_1, \dots, s_n) = \frac{\prod_i P(s_i)}{\prod_i Q(s_i)}$$

- Man berechnet in der Regel den Logarithmus:

$$LLR(S) = \log \left(\frac{\prod_i P(s_i)}{\prod_i Q(s_i)} \right) = \sum_{i=1}^n \log \frac{P(s_i)}{Q(s_i)}$$

Beispiel: $0.001 / 0.0002 = 5 \Rightarrow \log$ Likelihood Score 1.6

PWM Statistik

- Die einzelnen Summanden nennt man auch positionsspezifische Scores:

$$Score_i(s_i) = \log \frac{P(s_i)}{Q(s_i)}$$

- Diese werden in der Scoring-Matrix vorberechnet und zusammengefasst:

$$Score(S) = \sum_i Score_i$$

- Den Score für eine Sequenz kann man dann einfach durch ablesen und aufsummieren der Scores in der Matrix ausrechnen.

Weigth-Matrix						
	T	G	T	G	G	T
A	-1.0	-1.9	-1.2	-4.1	-2.5	-2.5
C	-2.5	-1.9	0.0	-4.1	-4.1	-1.2
G	-1.9	1.3	-2.5	1.4	1.3	-4.1
T	1.2	-2.5	1.0	-4.1	-2.5	1.3

Beispiel: **CGAGGT** $-2.5 + 1.3 - 1.2 + 1.4 + 1.3 + 1.3 = 1.7$

PWM Statistik

- Es kommt zu Problemen, wenn die Anzahl der bekannten Bindestellen gering ist, und einzelne Nukleotide an einer Position überhaupt nicht beobachtet werden.
- Beispiel: A an Position 4 wird 0 mal beobachtet.
Score= $\log(0/0.25)$ =nicht definiert
- Lösung: Es wird jeweils ein geringer „*Pseudocount*“ aufaddiert.
- Es gibt eine Reihe von Methoden basierend auf der Informationstheorie gute Pseudocounts auszuwählen.
- Beispiel: $f_{i,j} = \ln(((n_{i,j} + p_i) / (N+1)) / p_i)$

	T	G	T	G	G	T
A	5	2	4	0	1	1
C	1	2	14	0	0	4
G	2	52	1	57	55	0
T	49	1	38	0	1	52

Andere Maße: Entropie

- Anstelle der log-odds-Scores kann man auch ein anderes Maß benutzen: Die Entropie.
- Die Entropie gibt an, wie viel Information aus der Verteilung an einer bestimmten Stelle erhalten werden kann.
- Ist viel Information vorhanden, d.h. man ist sich sicher über die Identität eines Nukleotids, ist die Entropie klein.

Entropie

- Die Entropie wird in bits gemessen und gibt das Maß der Unsicherheit für die Identität eines Nukleotids an.
- Die Entropie für eine Position ist definiert als

$$H_c = - \sum_i p_{i,c} \log_2(p_{i,c})$$

und für eine Sequenz bzw. ein Alignment als

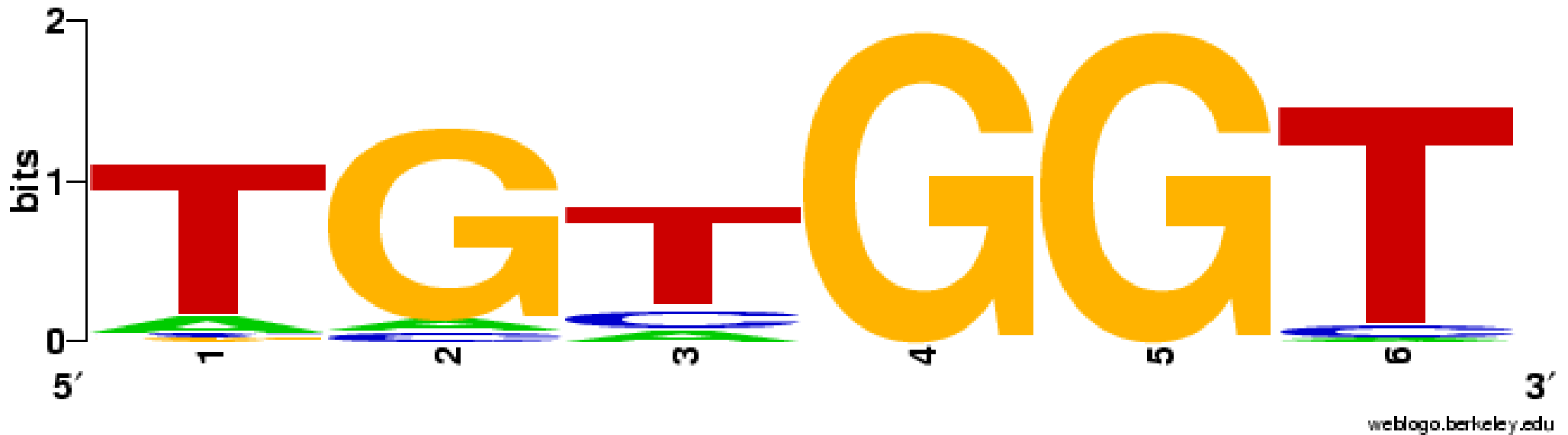
$$H = \sum_c H_c$$

Informationstheorie

- Wir stellen uns vor, wir hätten 64 umgedrehte Becher, und unter einem der Becher würde eine Kugel liegen. Was ist die minimale Anzahl an Fragen, die man stellen muß, um die Kugel sicher zu finden?
- Die minimale Anzahl ist 6: Die erste Frage wäre, ob die Kugel in der ersten Hälfte zu finden ist, dann in der Hälfte der Hälfte, usw.
- Insgesamt braucht man $\log_2(64)$ Fragen.
- Diesen Informationsgehalt bezeichnet man mit der Einheit bit. Der maximale Informationsgehalt bei 20 Aminosäuren ist $\log_2(20) = 4.32$ bits, bei Nukleotiden $\log_2(4) = 2$ bits.

Sequenzlogo

- Sequenzlogos werden benutzt um die Sequenzmuster aus PSSMs zu visualisieren.
- Die Gesamthöhe jeder Position ist proportional zum Informationsgehalt dieser Position.
- Die Höhe der Buchstaben ist proportional zum Anteil des Buchstaben an einer Position.



Konstruktion von Sequenzlogos

- Die Höhe der Buchstaben ist proportional zur *Reduktion* der Unsicherheit je Position; wenn die Entropie H_c ist, ist die Reduktion $H_{\max} - H_i$.
- Für Nukleinsäuren ist $H_{\max} = \log_2(4) = 2$ bits, für Proteine $H_{\max} = \log_2(20) = 4.32$ bits.
- Die Höhe der Buchstaben wird noch mit der relativen Häufigkeit des Nukleotids gewichtet, d.h.

$$\text{Höhe}_{i,c} = p_{i,c} (H_{\max} - p_{i,c} \log_2 p_{i,c})$$

Informationsgehalt (Stormo 1998)

- Der Informationsgehalt einer Matrix kann als mittlere Bindungsenergie eines TF an die Sequenz interpretiert werden.
- Wir nehmen an, daß alle Positionen unabhängig und gleichmäßig zu dieser Bindungsenergie beitragen.

$$I_{seq} = \sum_j \sum_b f_{b,j} \log \frac{f_{b,j}}{p_b}$$

Datenbanken

- TFBS
 - Transfac (<http://www.biobase.de>, neueste Versionen Komerziell)
 - TRRD (<http://www.msg.bionet.nsc.ru/mgs/gnw/trrd>)
- PSSMs
 - Transfac (<http://www.biobase.de>)
 - Jaspar (<http://jaspar.cgb.ki.se>)
 - TFD (<http://www.ifti.org>)
 - TESS (<http://www.cbil.upenn.edu/tess>)
- Bekannte Promotoren:
 - Eukariotic Promoter Database (EPD) – <http://www.epd.isb-sib.ch>

Verschiedene Matrix Suchprogramme

- Die meisten Suchprogramme für TFBS-Matrizen skalieren den log-odds Score zwischen 0 und 1.
- Matrix Search (Stormo 1995)

$$Score = \sum_{i=1}^L \log w_{i,b}$$

UND

$$Normalized\ Score = \frac{score}{score_{max}}$$

- Consite (Lenhard 2003)

$$Score = \sum_{i=1}^L \log w_{i,b} \quad ?$$

$$Normalized\ Score = 100 \frac{score - score_{min}}{score_{max} - score_{min}}$$

$$Current: \sum_{i=1}^L I(i) f_{i,b_i}$$

matrix
similarity
score

$$mSS = \frac{Current - Min}{Max - Min}$$

$$Min: \sum_{i=1}^L I(i) f_i^{min}$$

$$I(i) = \sum_{B \in (A,C,G,T)} f_{i,B} \ln(4f_{i,B})$$

$$Max: \sum_{i=1}^L I(i) f_i^{max}$$

Beispiel: Score für CGAGGT

$$\begin{array}{cccccc} & \text{Current} & & \text{Min} & \text{Max} & \text{Min} & \text{Score} \\ ((0.854*0.018)+(0.997*0.912)+(0.514*0.07)+(1.386*1)+(1.21*0.964)+(1.045*0.912)) & - & 0.141 & / & (5.493 & - & 0.141) = 0.808 \end{array}$$

Verschiedene Cutoffs für verschiedene PSSMs?

- Match (Transfac)
- Auswahl von Cutoffs (vorberechnete Listen):
 - Cut-offs, welche die Anzahl der Falsch Negativen minimieren (minFN):
Simulation ???
 - Cut-offs, welche die Falsch Positiven minimieren (minFP):
Anzahl der Hits in ($\sim 6 \times 10^6$ bp) 2. Exons
 - Cut-offs, welche die Summe aus beiden minimieren (minSum)
- Zwei Verschiedene Scores werden berechnet:
 - CORE (die ersten 5 am stärksten konservierten Positionen)
 - ganze MATRIX.

Sensitivität und Spezifität

- Man will Cutoffs finden, so daß man Idealerweise möglichst wenige wahre Bindestellen verliert (falsch Negative, FN) auf der anderen Seite aber auch nicht viele falsche Bindestellen findet (falsch Positive, FP). Entsprechend natürlich viele echt Positive (true positives, TP) findet.
- Sensitivität, wie hoch ist der Anteil der TPs die ich finde an allen wahren Positiven:

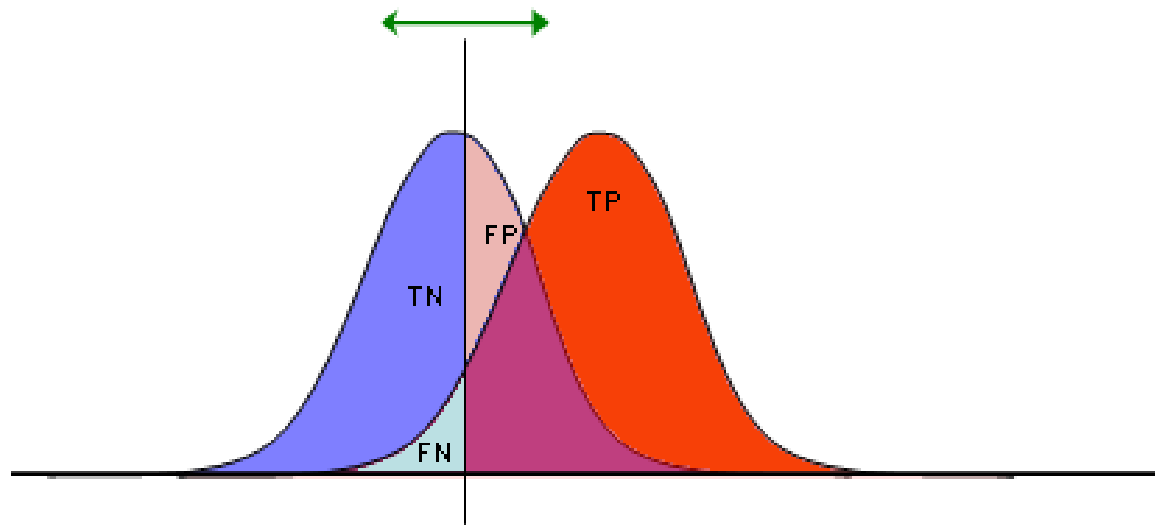
$$Sens = \frac{|TP|}{|TP| + |FN|}$$

- Spezifität, wie hoch ist der Anteil an FPs an allen ausgewählten Bindestellen:

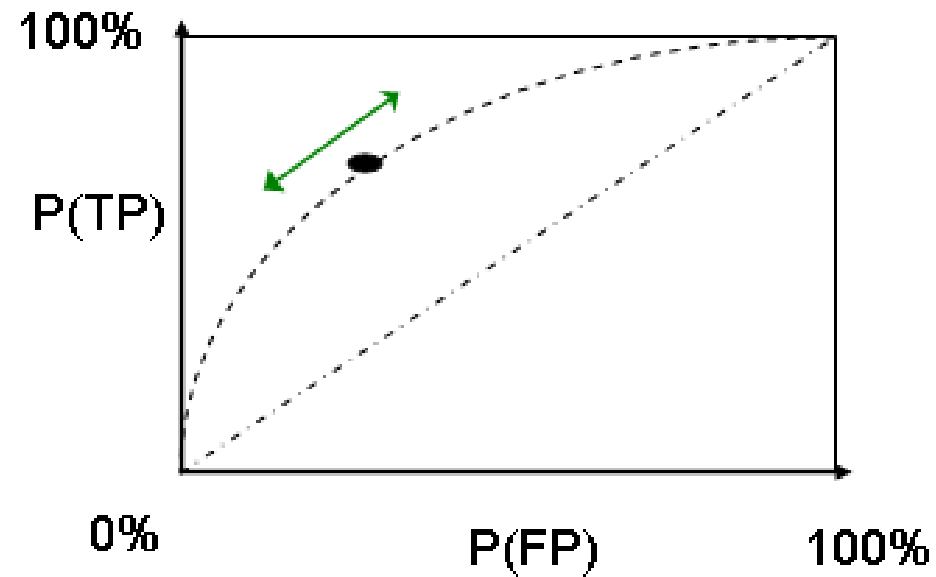
$$Spez = \frac{|FP|}{|FP| + |TP|}$$

- Bei TFBS-Suchen ist es oft unmöglich in beiden Werten gut zu sein.

Receiver Operating Characteristic (ROC)

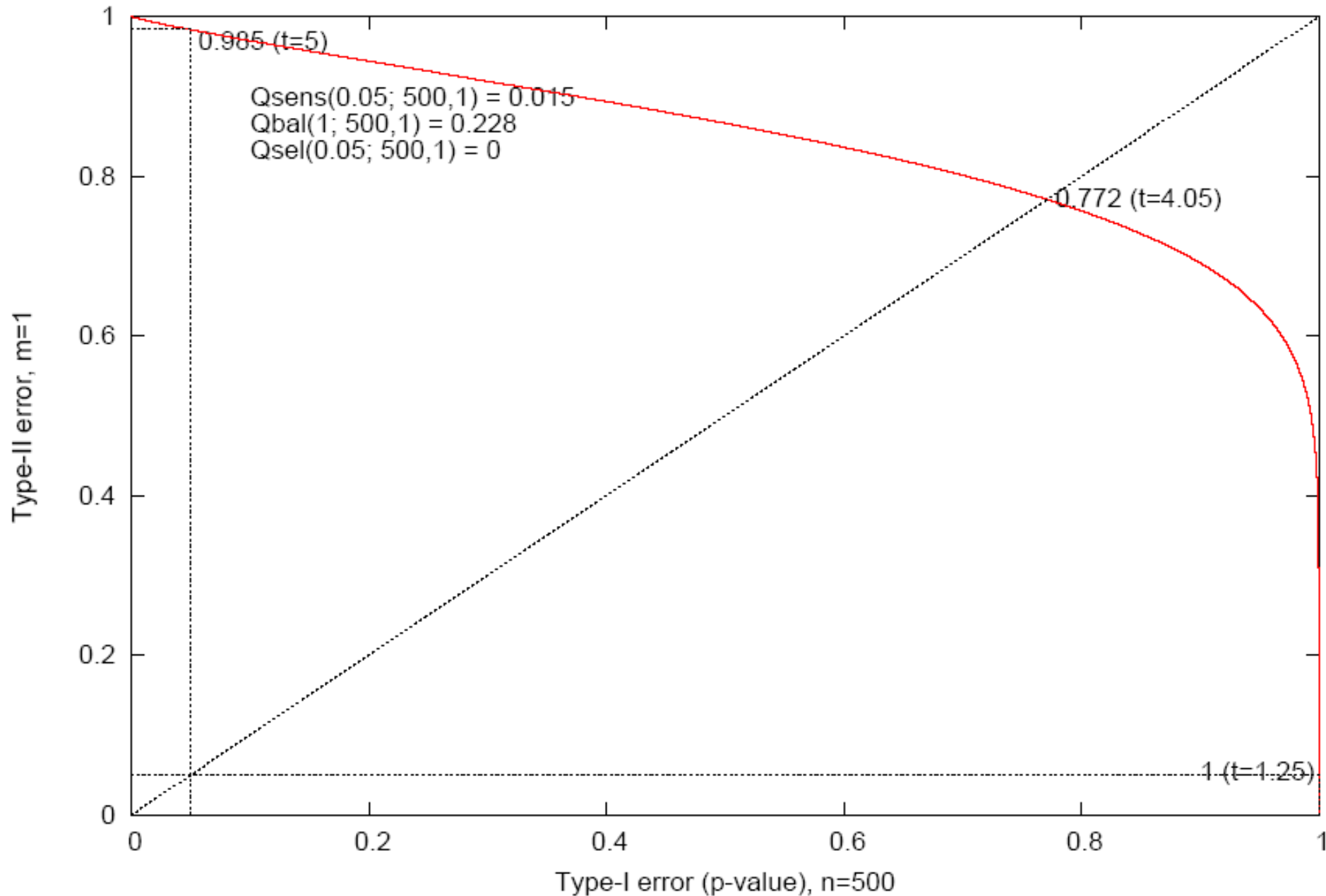


TP	FP
FN	TN
1	1



Receiver Operating Characteristic (ROC)

ROC Curve for Profile cap (V\$CAP-01), cap signal for transcription initiation



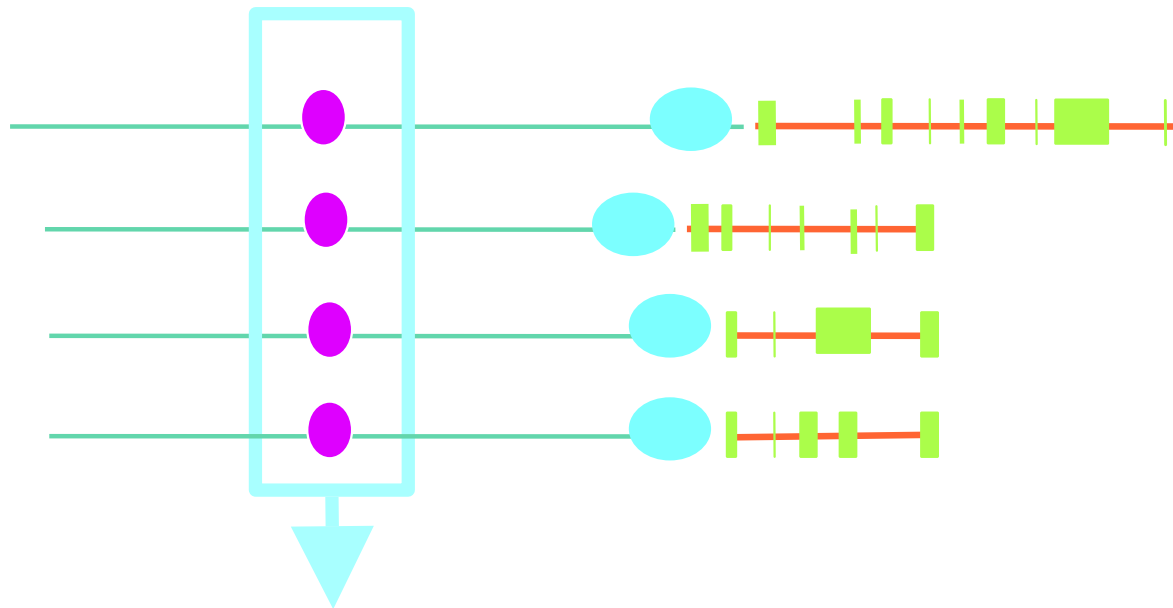
Methoden zum Finden von TFBS (PSSMs)

- Finden von TFBS durch gezielte Experimente in zu untersuchendem Promoter: Mutagenese, etc.
- Systematische experimentelle Suche nach Sequenzen, welche an einen TF binden.
- Chip-on-chip: Chromatin Immunoprecipitation, i.e. crosslinken von TFs an besetzte Bindestellen und testen, welche Sequenzen gebunden wurden über Microarray.
- Cluster von Genexpressionsdaten – Annahme, es gibt eine gemeinsame Bindestelle.

Problem: Gegeben ein Set an Sequenzen in denen bekanntermaßen eine gemeinsame TFBS vorkommt – finde die beste PSSM.

Finde die beste PSSM

- Annahme: Sequenzen erhalten gemeinsames Motif, z.B. Cluster aus Microarray, Chip-on-Chip Daten.



Position Specific Scoring Matrix (PSSM) oder Position Weight Matrix (PWM)

Experimentell bestimmte TFBS nicht aligniert.

ATAG**CTG**ACTA
 TGCA**CTG**AATTA
 TTCA**CTG**ACAG
 AGCT**CTG**AATG

	1	2	3	4	5	6	7	8	9	10	11
A	2	0	1	2	0	0	0	4	1	1	2
T	2	2	0	1	0	4	0	0	1	2	0
G	0	2	0	1	0	0	4	0	0	0	2
C	0	0	3	0	4	0	0	0	2	0	0

Methoden

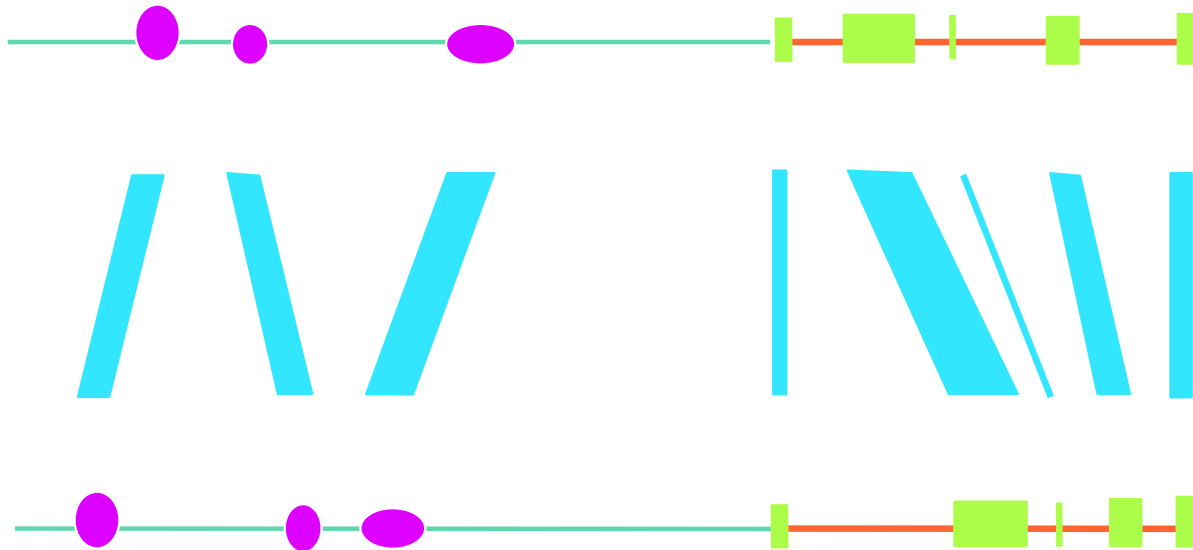
- MEME (EM algorithmus, Bailey/Elkan, *Machine Learning*, 1995)
- Gibbs Sampler (Lawrence/Altschul/Boguski/Liu, *Science*, 1993)
- AlignACE (Roth/Hughes/Estep/Church, *Nature Biotechnology*, 1998)

Methoden die Spezifität zu erhöhen

- Einschränken der Suche auf TFBS-Bindestellen, welche zwischen mehreren Spezies konserviert sind – phylogenetic Footprinting.
- Suchen nach kombinationen von TFBS:
 - Gezielte Kombinationen – Transcriptional Modules (Werner, 2001)
 - Kombinieren von P-values mehrerer benachbarter TFBS. (Johansson et al, Bioinformatics, 2003)

Sequenzähnlichkeit zwischen Orthologen Genen kann benutzt werden um funktionelle Elemente zu identifizieren.

Menschliches Gen



Sequenz Ähnlichkeit zwischen Orthologen ist weitestgehend auf funktionelle Einheiten beschränkt.

Exons oder regulatorische Elemente.

Orthologes Maus Gen

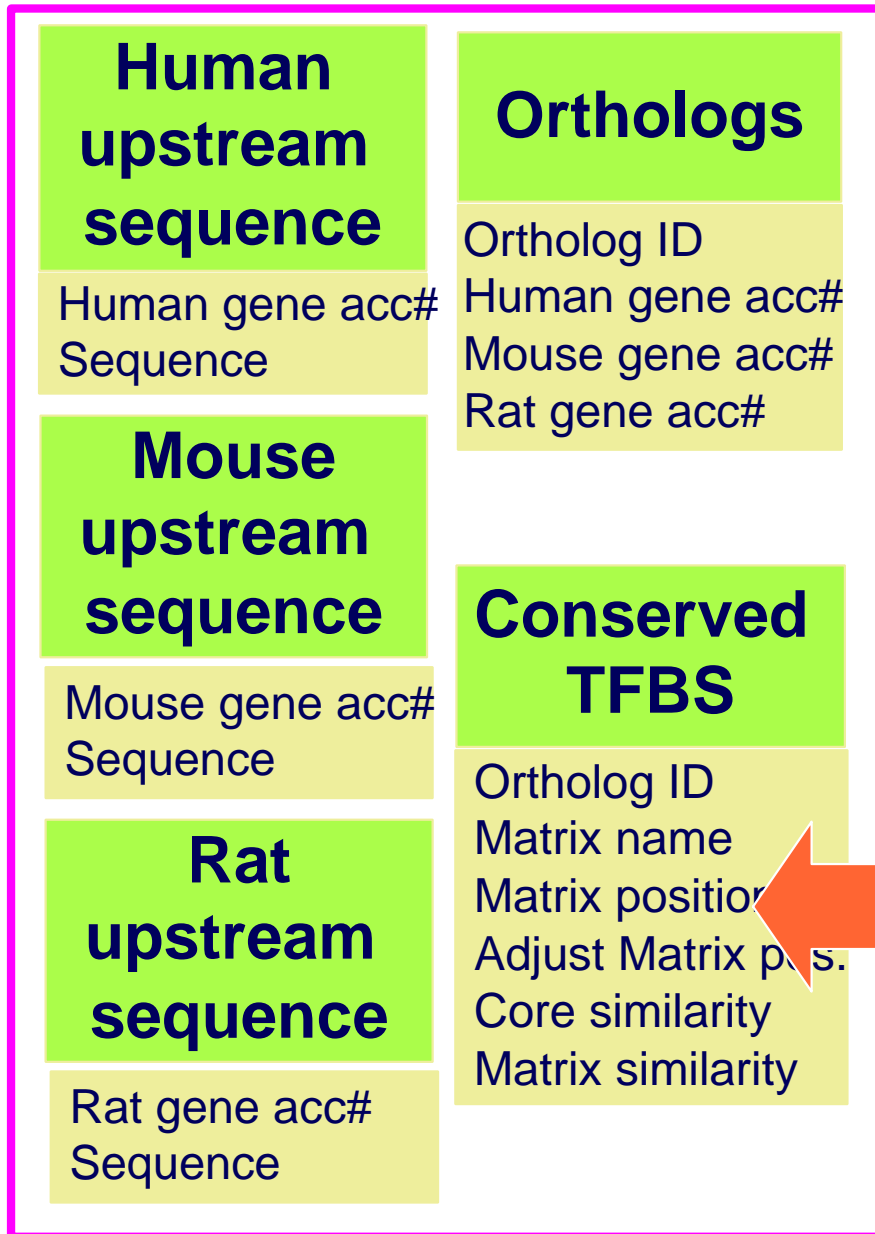
Tools Datenbanken zum finden konservierter TFBS

- RVISTA (<http://rvista.dcode.org/>)
- CONSITE (<http://www.phylofoot.org/>)
- CORG (<http://corg.molgen.mpg.de/>)

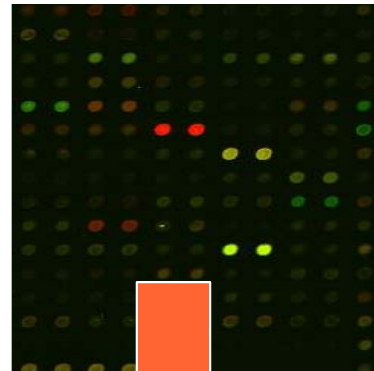
Konservierte regulatorisch Elemente

- Finde die Upstream Bereiche von Genen im Genom
Problem: Transcription Start Site nicht immer bekannt. Wie groß soll der Bereich sein, den man durchsucht?
- Aligniere genomische Sequenzen oder finde hoch konservierte Regionen.
Genom Alignierer: MAVID, MLAGAN, MUSCLE, etc.
Problem: Große Genomische Bereiche. Finde stark konservierte Cores, Repeats, Inversionen.
- Finde TFBS
- Vergleiche Konservierungsprofil mit Positionen der TFBS.

Datenbank abfragen



cDNA microarrays



AA025... in kinase C
AA0342... egral membra
AA0356... nction plakogl
AA0357... 83' exoribonuc
AI87009... LKL motif kina:
AA0399... omboxane A2
AI31573... D28 antigen (T
AI3177... bby like protei
AI3177... egrin, beta 72
AI3177... nc finger prote
AI33060... meo box C6
AI39467... allikrein 12
AI86519... egrin, beta 7"
AI87240... efoldin 1

- Datenbank Kann wie mit einer Liste von Genen aus einem Microarray abgefragt werden.
- Statistik ähnlich wie bei „Gene Set Enrichment Analysis“ bzw. GO Statistik Kann gemacht werden.

Acknowledgements – Slides geborgt von

- Terry Speed
- Benedikt Brors