

On the Power of Profiles
for
Transcription Factor Binding Site (TFBS)
Detection

(S. Rahmann, T. Müller, M. Vingron)

Übersicht

- TFBS finden (gene finding & regulation)
- Matrizen zur Beschreibung von TFBS
- Regularisierung
- Beurteilen der Profilqualität
- Suche eines optimalen Score Thresholds
- Ergebnisse

Matrizen

- *count matrix* C mit Einträgen C_{ij}

- Spaltensumme N_i

- *profile matrix* P mit Einträgen $P_{ij} = C_{ij}/N_i$

- Spaltensumme 1

- *score matrix* S mit Einträgen S_{ij}

$j \backslash i$	1	2	3	4	5	...	I
A							
C							
G							
T							
Σ							

Regularisierung

- Problem: wenig Daten
 - Overfitting
 - Nullen in der count matrix
- Idee: Regularisierung
 - Addieren eines kleinen Werts auf die Einträge der count matrix

Regularisierung

- gute Regularisierungsmethode: nicht trivial
 - signifikante Signale nicht stören
 - keine nicht signifikanten Signale erzeugen

Regularisierung

- regularisierte Matrix $C' = C + R$
- Matrix R mit „pseudocounts“; Spaltensumme W_i
→ profile matrix:
$$P'_{ij} = \frac{C_{ij} + R_{ij}}{N_i + W_i}$$
- $w = \frac{W_i}{N_i + W_i}$ Maß für den Einfluß der pseudocounts
- optimale Wahl von w

Regularisierung nach Rahmann et al.

- Gesamt-Nucleotid-Zusammensetzung nicht verändern
- Regularisierung abhängig von Signalstärke

Regularisierung nach Rahmann et al.

- Fokus auf eine Spalte i
- τ ist die Spalte P_i der Profilmatrix:
 $\tau_j = C_{ij}/N_i$
- definieren einer „*regularizing distribution*“ ρ und eines Gewichtungsfaktors

- $C'_i = N_i \cdot \tau + W_i \cdot \rho$ $w = \frac{W_i}{N_i + W_i}$

- $P'_i = \delta(w) = (1-w) \cdot \tau + w \cdot \rho$

Bestimmung von ρ

- keine Veränderung der Gesamt-Nukleotid-Zusammensetzung

$$\rightarrow \rho_j = \frac{\sum_i C_{ij}}{\sum_i N_i}$$

- Annahme: $\rho_j > 0 \quad \forall j$

- wenn nicht:

$$\rho_j = \frac{1/4 + \sum_i C_{ij}}{1 + \sum_i N_i}$$

Bestimmung von w

- $P'_i = (1-w) \cdot \tau + w \cdot \rho = \delta(w)$
- $\delta(0) = \tau \quad \delta(1) = \rho$
- definieren der skalierten relativen Entropie $\Delta(w)$ zw. $\delta(w)$ u. $\delta(1) = \rho$:

$$\Delta(w) = 2 N_i \sum_j \delta(w)_j \cdot \ln \left(\frac{\delta(w)_j}{\rho_j} \right)$$

- Erwartungswert E von der relativen Entropie von N Samples aus ρ und ρ selbst

Bestimmung von w

- Finde ein w für das gilt:

$$\Delta(w) = \Delta(0) - E$$

- wenn $\Delta(0) = E \Rightarrow \Delta(w) = 0 \Rightarrow w = 1$

(τ wird nicht berücksichtigt, ρ voll)

- falls $\Delta(0) < E \Rightarrow w := 1$

- in der Praxis: heuristischer Wert $E = 1,5$

- Score-Matrix S mit Einträgen $S_{ij} = \log (P_{ij}/\pi_j)$
- Wahrscheinlichkeitsmodell für den Hintergrund:
 - i.i.d. Modell (independent, identically distributed)
 - Profilmatrix Π
 - jede Spalte besteht aus dem gleichen Wkheitsvektor π

PSSMs

- $\underline{S} := \min_{W \in \Sigma^L} \text{Score}(W) = \sum_{i=1}^L \min_{j \in \Sigma} S_{ij}$; kleinstmöglicher Score
- $\overline{S} := \max_{W \in \Sigma^L} \text{Score}(W) = \sum_{i=1}^L \max_{j \in \Sigma} S_{ij}$; größtmöglicher Score

Fehlerwahrscheinlichkeiten

- Type-I window error probability $\alpha(t)$:
 - Wkeit, einen Score $\geq t$ zu beobachten, obwohl das Sequenzfenster W durch das Hintergrundmodell erzeugt wird (False Positive)
- Type-I sequence error probability $\alpha_n(t)$:
 - Wkeit dafür, dass mindestens eines von n aufeinanderfolgenden überlappenden Sequenzfenstern einen Score $\geq t$ erreicht
 - (Annahme: die ganze Sequenz wird durch Hintergrundmodell erzeugt)

Fehlerwahrscheinlichkeiten

- Ein Schwellenwert t mit $\alpha_n(t) \leq 0,05$ bis $0,01$ wird als signifikant angesehen
- Für einen best. Score s wird
 - $\alpha(s)$ auch window p-value von s genannt
 - $\alpha_n(s)$ sequence p-value von s genannt

Fehlerwahrscheinlichkeiten

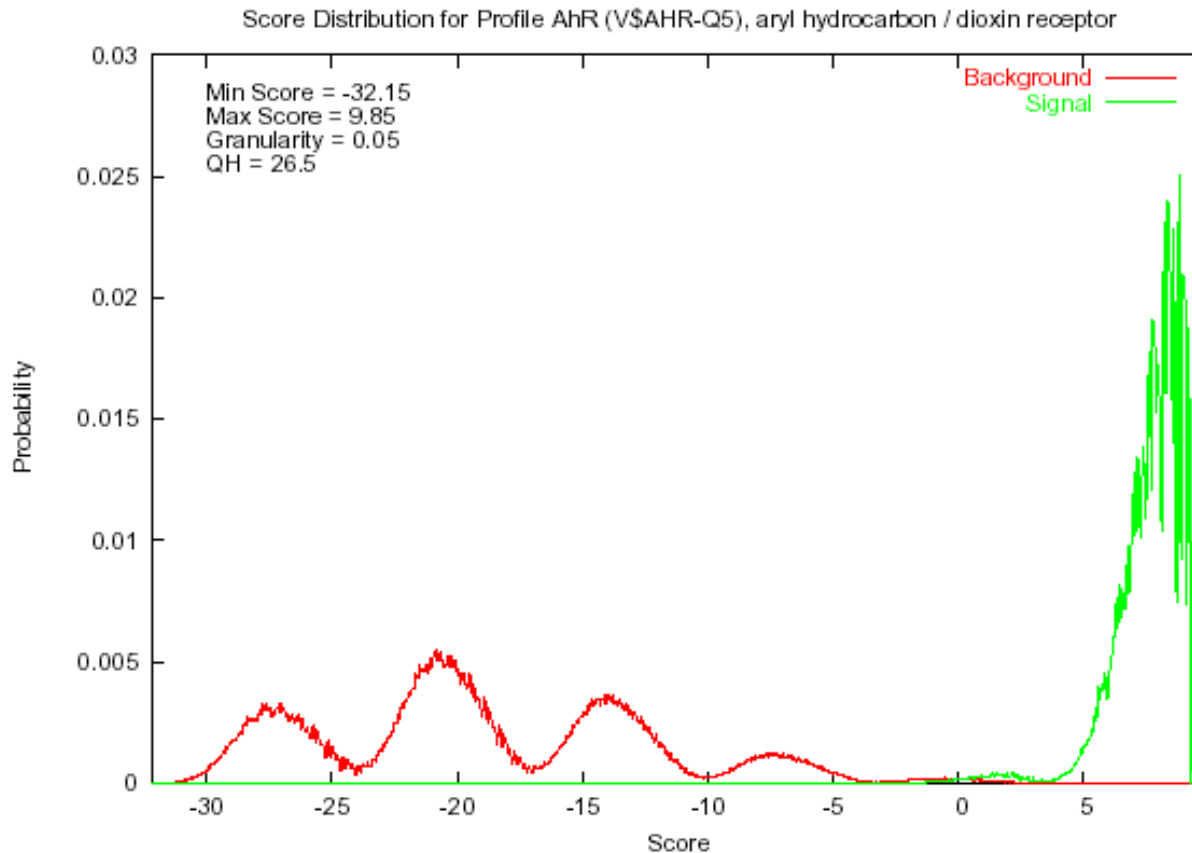
- **type-II error probability $\beta(t)$:**
 - Wkeit, einen Score $< t$ zu erhalten, obwohl das Sequenzfenster durch das Signalprofil erzeugt wird (False negative)
- eine type-II sequence error probability macht keinen Sinn (die ganze Sequenz kann nicht durch ein und dasselbe Profil erzeugt werden)
- **m-instance type-II error probability $\beta_m(t)$:**
 - Wkeit, dass mindestens 1 von m unabhängigen Fenstern einen Score $< t$ erhält
 - (Annahme: alle Fenster wurden durch das Signalprofil erzeugt)
- **power: $1-\beta_m(t)$**

Fehlerwahrscheinlichkeiten

- hohe Signifikanz = kleiner p-value
 - die Wkeit eines Typ-I-Fehlers (FP) ist klein
- keine Information über power
- hoher threshold:
 - zwar wenig falsch positive
 - aber auch wenig richtig positive (viel falsch negative!)

Die genaue Verteilung des Scores

- oft wird eine Gauß'sche Verteilung angenommen

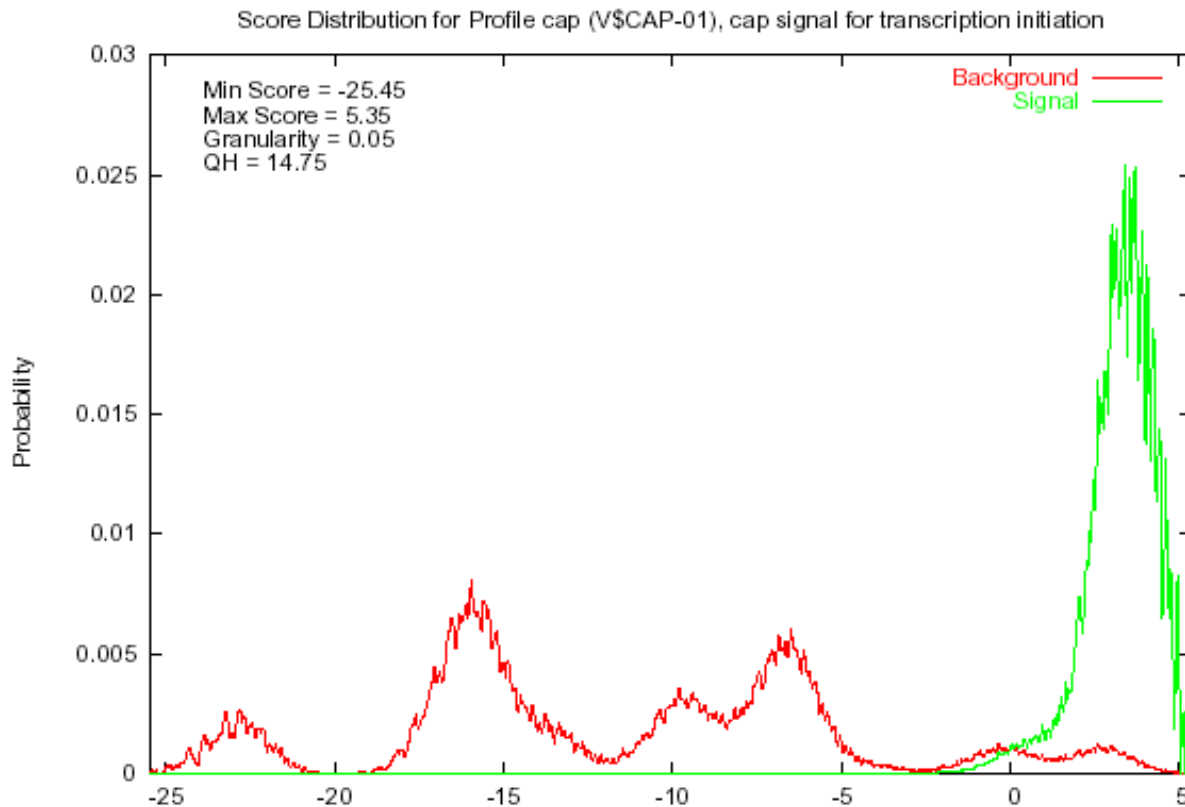


Die genaue Verteilung des Scores

- Zufallsvariable $X := \text{Score}(W)$; $X = \sum_{i=1}^L X_i$
- partielle Summen: $X^k := \sum_{i=1}^k X_i$
 - Score bis zur Stelle k
- X_i sind unabhängig
- $X^{k+1} = X^k + X_{k+1}$
- $\mathbb{P}(X^{k+1} = x) = \mathbb{P}(X^k + X_{k+1} = x)$
 $= \int \mathbb{P}(X^k = z) \cdot \mathbb{P}(X_{k+1} = x-z) dz$ (Faltung)

Berechnen der window error probabilities $\alpha(t)$ und $\beta(t)$

- $\alpha(t) := \mathbb{P}_{\Pi}(X \geq t) = \int_t^{\bar{s}} \mathbb{P}_{\Pi}(X=s) ds$ $\beta(t) := \mathbb{P}_P(X < t) = \int_{\underline{s}}^t \mathbb{P}_P(X=s) ds$



- exakte Berechnung schwierig (Überlappung => aufeinanderfolgende Scores X_i nicht unabhängig)
- Annahme: gesamte Sequenz wird von Π erzeugt
- $\alpha_n(t) = \mathbb{P}(X(i) \geq t \text{ for at least one } i \in \{1, \dots, n\})$

$$\alpha_n(t) = 1 - \mathbb{P}(X(t) < t \text{ for all } i \in \{1, \dots, n\})$$

$$\alpha_n(t) \approx 1 - (1 - \alpha(t))^n \approx 1 - e^{-n\alpha(t)}$$

Approximation gültig für große n und $n\alpha(t) \ll 1$

- in Realität: m Instanzen eines Signals
- $\beta_m(t)$: mindestens 1 Instanz hat einen Score $< t$
- Annahme: echte TFBS sind selten \Rightarrow keine Überlappungen
 \Rightarrow Unabhängigkeit
- $X(i)$: Score der i-ten Instanz
- $\beta_m(t) = \mathbb{P}_p(X(i) < t \text{ for at least one } i) = 1 - (1 - \beta(t))^m$

Beurteilen der Profilqualität

- Ziel: quantifizieren, wie gut ein Signalprofil P von einem Hintergrund Π getrennt ist
- Differenz der erwarteten Scores (Q_H)
- Sensitivität (Q_{sens})
- Selektivität (Q_{sel})
- Error balance (Q_{bal})
- hohes Q -Maß = gute Trennung

Differenz der erwarteten Scores

- Erwartungswerte der Scores $E_P[X]$ und $E_{\Pi}[X]$
 - Differenz: Maß für Profilqualität
 - $E_P[X_i] = \sum_{j \in \Sigma} P_{ij} S_{ij}$
 - $E_P[X] = \sum_{i=1}^L E_P[X_i]$

Differenz der erwarteten Scores

$$E_P [X_i] = \sum_{j \in \Sigma} P_{ij} S_{ij} = \sum_{j \in \Sigma} P_{ij} \log(P_{ij} / \pi_j)$$

$$E_P [X_i] = H(P_i \| \pi)$$

relative Entropie (immer positiv)

$$E_P [X] = \sum_{i=1}^L E_P [X_i] = H(P \| \Pi) \geq 0$$

=> für $W \sim P$ wird ein positiver Score erwartet

Differenz der erwarteten Scores

$$E_{\Pi} [X_i] = \sum_{j \in \Sigma} \pi_j S_{ij} = \sum_{j \in \Sigma} \pi_j \log(P_{ij} / \pi_j) = - \sum_{j \in \Sigma} \pi_j \log(\pi_j / P_{ij})$$

$$E_{\Pi} [X_i] = -H(\pi \parallel P_i)$$

$$E_{\Pi} [X] = -H(\Pi \parallel P) \leq 0$$

- für $W \sim \Pi$ wird ein negativer Score erwartet

Differenz der erwarteten Scores

- leicht zu berechnen
- $Q_H := E_P[X] - E_\Pi[X]$
 $= H(P||\Pi) + H(\Pi||P)$

Sensitivität

- finde ich auch alle TFBS, die es gibt?
- (Anzahl n der Sequenzfenster der Länge L) ≥ 1
- Instance number $m \geq 1$
- $Q_{\text{sens}}(\alpha, n, m) := 1 - \beta_m(t)$
- Fähigkeit, bei einer bestimmten Signifikanz α einen TP zu erkennen

Selektivität

- sind die gefundenen auch echte TFBS?
- $Q_{\text{sel}}(\beta, m, n) := 1 - \alpha_n(t)$

Error balance

- plot von Type-II gegen Type-I error probability
- $\beta_m(t)$ gegen $\alpha_n(t)$
- **ROC-Kurve** (receiver operator characteristic)
- wir interessieren uns für den Punkt der Kurve, wo

$$\alpha_n(t) = \beta_m(t) \quad \text{bzw.}$$

$$\beta_m = c\alpha_n \quad (\alpha_n \text{ ist } c \text{ mal wichtiger als } \beta_m)$$

- $Q_{\text{bal}}(c,n,m) := 1 - \beta_m(t)$
mit t so dass $\beta_m(t) = c \alpha_n(t)$

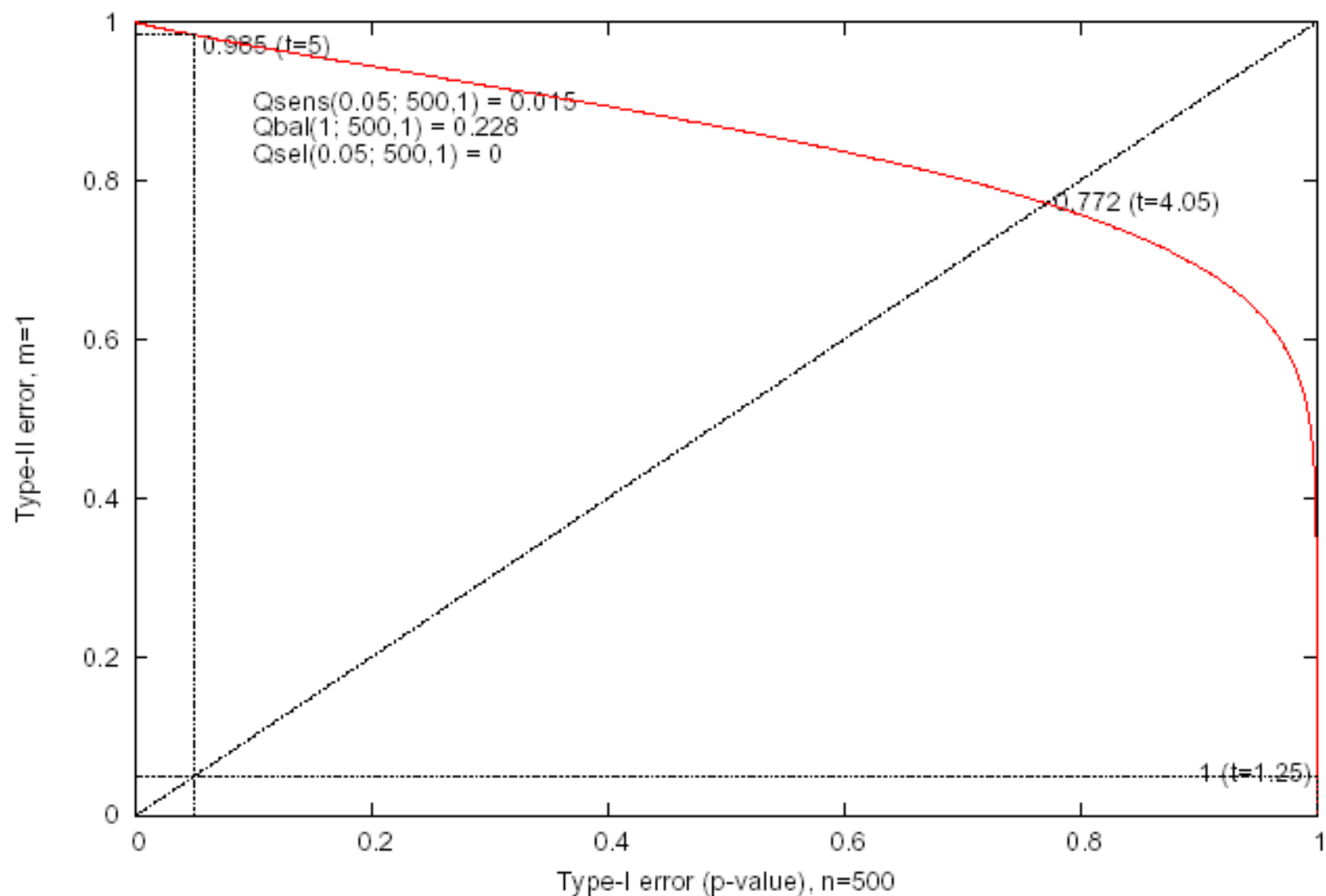


Figure 2: The ROC curve shows how type-II error β_m varies with type-I error α_n . The dotted lines indicate the value of β_m where $\alpha_n = 0.05$, the value of α_n where $\beta_m = 0.05$ and the point where $\alpha_n = \beta_m$.

Suche eines optimalen Score Thresholds

- Üblich ist:
 - t so, dass p-value der Sequenz $\alpha_n(t)$ einen festen Wert hat (z.B. 0,05 für Sequenzlänge 500)
 - dabei wird power ($Q_{\text{sens}}(\alpha, n, 1)$) nicht berücksichtigt
- => Fehler ausbalancieren (ROC-Kurve)

Ergebnisse in TRANSFAC

- bei hoher Selektivität, niedrige Sensitivität und umgekehrt
- beim Betrachten eines der beiden Kriterien:
nur ca. 20% sind richtig gut ($Q=0,95$)

Qualität des Verfahrens

- Hat die Regularisierung was gebracht?
 - simulierte Matrix P („echtes“ Profil)
 - N Samples aus P ziehen => count matrix
 - count matrix regularisieren und entsprechendes Profil \hat{P} berechnen
 - Berechnen des Abstands von Qualitätsmaßen zw. „echt“ & „samples“:
 - $Q_H = E_P[X] - E_\pi[X]$
 - $\hat{Q}_H = E_P[\hat{X}] - E_\pi[\hat{X}]$
 - $|Q_H - \hat{Q}_H|$
 - leicht besser als vergleichbare Methoden