



Johann Wolfgang Goethe-Universität  
Frankfurt am Main

---

# Vorhersage von Protein-Funktionen

Patrick Pfeffer

# Überblick

- Motivation
- Einleitung
- Methode
- Markov Random Fields
- Der Gibbs Sampler
- Parameter-Schätzung
- Bayes'sche Analyse
- Resultate

# Motivation

- Es sollen mittels Bayes'scher Methoden Vorhersagen über Proteinfunktionen gemacht werden
- Die Wahrscheinlichkeit ist dann eine handfeste Aussage, inwiefern eine potentielle Funktion in Erscheinung treten kann
- Gearbeitet wird mit Hefe Proteinen aus der YPD ([www.incyte.com](http://www.incyte.com))
- Die Protein-Protein Interaction Data wird aus dem Münchner Informations Center für Protein Sequenzen bezogen (MIPS, [mips.gsf.de](http://mips.gsf.de))

# Einleitung

- Die größte Herausforderung sind die Funktionen unbekannter Proteine herauszufinden
- Der Klassische Weg ist, Homologe zu finden
  - Vorhersage der Funktion basierend auf Sequenzhomologien; Mustererkennung bei beiden Sequenzen
- Proteine haben eine mächtige Stellung in zellulären Vorgängen und etliche dieser hängen von Protein-Protein Interaktionen ab
- Aussagen über Funktion aufgrund seines Reaktionspartners
- Die Interaktionspartner können trotzdem unterschiedlicher funktioneller Gruppen angehören

# Methode

- Das Protein-Protein Interaktions Netzwerk soll eine Nachbarschaftsstruktur entlang aller Proteine beschreiben
- Der Nachbar hat dann Information über die Funktion des unbekanntes Proteins
- Wir wollen Proteine mit einer Wahrscheinlichkeit behaften, dass diese die Funktion aufweisen
- Es gibt mehrere Konfigurationen:
  - Nur ein Partner hat die Funktion, beide, keiner
- Ein „Belief Network“ → was zu der Theorie von Markov Random Fields führt
- Die Gewichtung der Parameter und die Bestimmung der  $W$ 'keiten wird der zentrale Punkt sein
- Durch biologische Experimente ist bekannt, welches Protein mit welchem eine Reaktion eingeht

# Methode

- Ein Genom hat  $N$  Proteine,  $P_1 \dots P_N$ ,  $M$  funktionelle Gruppen  $F_1 \dots F_M$ , nicht studierte Proteine  $P_1 \dots P_n$  und studierte sind  $P_{n+1} \dots P_{n+m}$  mit  $N = n+m$
- Es werden die einzelnen funktionellen Gruppen eines Proteins separat betrachtet
- Sei  $X_i = 1$  bei einer vorhandenen Funktion, sonst 0
- Sei  $X = (X_1 \dots X_{n+m})$  die Konfiguration der Funktionsbezeichnung (0 oder 1)
  - Wobei  $X_1 = \lambda_1, \dots, X_n = \lambda_n$  nicht studiert
  - $X_{n+1} = \mu_1, \dots, X_{n+m} = \mu_m$  studiert

$$o_{ij} = \begin{cases} 1 & \text{if } P_i \text{ and } P_j \text{ are observed to interact} \\ 0 & \text{otherwise.} \end{cases}$$

$$S = \{P_i \leftrightarrow P_j : o_{ij} = 1, i, j = 1 \dots$$

$N\}$

# Methode

- Sei  $S$  die Menge aller Interaktions-Partner
- Es wird als Graph aufgefasst
  - $P_i$  = das  $i$ -te Protein,  $i=1,2,\dots,N$
  - $Nei(i)$ : Nachbarn des  $P_i$ . Die Menge aller Proteine, die mit  $P_i$  interagieren
  - $F_j$  = die  $j$ -te Funktionskategorie,  $j=1,2,\dots,M$
  - $\pi_j$  = Der Teil aller Proteine, der diese Funktion  $F_j$  hat

# Markov Random Fields (MRF)

- Bedingt durch die Funktion des studierten Proteins wollen wir die  $W$ 'keit errechnen, dass ein unstudiertes Protein diese Funktion besitzt ( $\pi$ )
- Ohne das Interaktions-Netzwerk zu betrachten, ist die  $W$ 'keit einer Konfiguration  $X$  proportional zu:

$$\prod_{i=1}^N \pi^{x_i} (1 - \pi)^{1-x_i} = \left( \frac{\pi}{1 - \pi} \right)^{N_1} (1 - \pi)^N$$

- Jetzt wird das Interaktion-Netzwerk betrachtet unter der Tatsache, dass die  $W$ 'keit für gleiche Funktionen interagierender Partner höher ist als umgekehrt

# Markov Random Fields (MRF)

- Unter Bedingung der eben errechneten Funktionsbezeichnung ist die  $W$ 'keit des Netzwerkes proportional zu:

$$\exp(\beta N_{01} + \gamma N_{11} + N_{00})$$

$$\begin{aligned} N_{11} &= \sum_{(i,j) \in S} x_i x_j \\ &= \#\{(1 \leftrightarrow 1) \text{ pairs in } S\}, \end{aligned}$$

$$\begin{aligned} N_{10} &= \sum_{(i,j) \in S} (1 - x_i) x_j + (1 - x_j) x_i \\ &= \#\{(1 \leftrightarrow 0) \text{ pairs in } S\}, \end{aligned}$$

$$\begin{aligned} N_{00} &= \sum_{(i,j) \in S} (1 - x_i)(1 - x_j) \\ &= \#\{(0 \leftrightarrow 0) \text{ pairs in } S\}. \end{aligned}$$

# Markov Random Fields (MRF)

- Die totale W'keit der Funktionsbezeichnung ist proportional zu  $\exp(-U(x))$ :

$$U(X) = -\alpha N_1 - \beta N_{10} - \gamma N_{11} - N_{00}$$

$$\begin{aligned} &= -\alpha \sum_{i=1}^N x_i - \beta \sum_{(i,j) \in S} x_i x_j \\ &\quad - \gamma \sum_{(i,j) \in S} (1 - x_i) x_j + (1 - x_j) x_i \\ &\quad - \sum_{(i,j) \in S} (1 - x_i)(1 - x_j). \end{aligned}$$

$$\alpha = \log\left(\frac{\pi}{1-\pi}\right)$$

- $U(x)$  sei die Potential Funktion, welche eine globale Gibbs-Verteilung des Netzwerks definiert

# Markov Random Fields (MRF)

- $\Theta$  sind die Parameter der einzelnen  $W$ 'keiten mit  $\Theta = (\alpha, \beta, \gamma)$
- Die Gibbs Verteilung gibt die  $W$ 'keit einer Funktion an unter der Bedingung eines Parameters
- In der Gibbs-Verteilung ist  $Z(\Theta)$  im Nenner, weil es hier um eine  $W$ 'keit geht, weiter oben ist z.B. der Parameter  $\beta$  nicht im Nenner, weil es um die Summe aller Proteine geht
- Die Partitionsfunktion  $Z$ , die den Parameterwert angibt, ist eine Summe über alle vorhandenen Funktionsbezeichnungen, eine Natürliche Zahl mit positivem Wert oder 0, da die Summanden alle 0 oder 1 sind

$$\Pr(X | \theta) = \frac{1}{Z(\theta)} \exp(-U(x)) \quad Z(\theta) = \sum_x \exp(-U(x))$$

# Markov Random Fields (MRF)

- Die Gibbs Verteilung gibt die Funktionsbezeichnung aller Proteine an, wir haben aber nur Daten für jene
- Mit einem Bayes'schen Ansatz soll die spätere Verteilung von  $(X_1, \dots, X_n)$  mit den gegebenen Daten gefunden werden ( $n$ = das letzte unstudierte Protein)
- Finden der Funktionsbezeichnung unter Bedingung der gegebenen Daten

$$\Pr(X_1, \dots, X_n \mid X_{n+1} = \mu_1, \dots, X_{n+m} = \mu_m)$$

- Jetzt noch über alle möglichen Konfigurationen  $X_j, j \neq i, 1 \leq j \leq n$  summieren mithilfe des Gibbs Samplers

# Der Gibbs Sampler

- Wenn alle Funktionen des Interaktionspartners gegeben sind, kann die  $W$ 'keit für die Funktion ermittelt werden
- Hier gehen wir von gegebenen Parametern aus
- Wird die Prozedur wiederholt, kann die Funktionsbezeichnung für alle unstudierten Proteine gewonnen werden

# Der Gibbs Sampler

*1+ = die Summe aller  
W'keiten*

$$\begin{aligned} & \Pr(X_i = 1 \mid X_{[-i]}, \theta) \\ &= \frac{\Pr((X_i = 1, X_{[-i]}) \mid \theta)}{\Pr((X_i = 1, X_{[-i]}) \mid \theta) + \Pr((X_i = 0, X_{[-i]}) \mid \theta)} \\ &= \frac{e^{\alpha + (\beta - 1)M_0^{(i)} + (\gamma - \beta)M_1^{(i)}}}{1 + e^{\alpha + (\beta - 1)M_0^{(i)} + (\gamma - \beta)M_1^{(i)}}}, \end{aligned}$$

$$X_{[-i]} = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_{n+m})$$

$$M_0^{(i)} = \#\{j \in \text{Nei}(i) : X_j = 0\}, M_1^{(i)} = \#\{j \in \text{Nei}(i) : X_j = 1\}$$

# Parameter-Schätzung

- Wir betrachten das Subnetzwerk der bekannten Proteine und schätzen die Parameter

$$S' = \{P_i <- - > P_j : o_{ij} = 1, \quad i, j = n+1, \dots, n+m\}$$

- Es wird ein quasi-Likelihood Ansatz aus der Bild Analyse benutzt

$$\log \frac{\Pr(X_i = 1 | X_{[-i]}, \theta)}{1.0 - \Pr(X_i = 1 | X_{[-i]}, \theta)} \quad \text{W'keit durch Gegen-W'keit}$$
$$= \alpha + (\beta - 1)M_0^{(i)} + (\gamma - \beta)M_1^{(i)}$$

- Die Funktionsbezeichnung ist nicht unabhängig, doch (2) ist bereits eine Parameterfunktion, deshalb wird ein quasi-Likelihood Ansatz verwendet

# Bayes'sche Analyse

- 1.: Wir schätzen die W'keit  $\pi$  das ein Protein eine Funktion hat (ohne die Information des Interaktions Netzwerkes) mit dem Anteil an Proteinen, die diese Funktion haben
- 2.: Parameterschätzung mit dem quasi-Likelihood Ansatz
- Jetzt haben wir folgenden Algorithmus:
  - Random set value of missing data  $X_i = \lambda_i, i=1 \dots n$  mit W'keit  $\pi$
  - For each Protein  $P_i$  using (3), update value of  $X_i$
  - Repeat 2 until all posterior propabilities  $\Pr(X_i | X_{[-i]})$  are stabilized
- Beim Gibbs Sampling wird zwischen der *burn-in*-Periode und der *lag*-Periode unterschieden

# Bayes'sche Analyse

- *Burn-in*:
  - die Zeit, bis sich der Markov'sche Prozess stabilisiert hat, also bis die Initialwerte reduziert bzw. eliminiert worden sind
- Danach wird  $W$ 'keit approximiert, dass ein Protein eine bestimmte Funktion hat durch Mittelwertbildung der Simulationsresultate (lag-Periode), um die Abhängigkeit des Markov'schen Prozesses zu reduzieren oder eliminieren
- Burn = 100 Schritte, lag = 10 Schritte als Beispiel
- Simulationsschritte sind insgesamt 2000
- Dieser Prozess wird für jede Funktionelle Gruppe eines Proteins wiederholt und die  $W$ 'keit für die Funktion eines Proteins bestimmt

# Resultate

- Gefolgert werden sollen unstudierte Hefe Protein-Funktionen durch Benutzung von Daten von studierten Proteinen von der YPD
- YPD hat 6416 Proteine basierend auf den 3 Kriterien ganz oben genannt
- Für Protein Interaktionen wurden Daten der MIPS benutzt
- MIPS hat 2439 Interaktionspartner (120 selbst Partner ausgeschlossen)
- Die Durchschnittliche # an Interaktionspartnern beträgt 2.6

# Resultate

- Anteilig:
  - Biochemische Funktionen: 79%
  - Subzelluläre Lokation: 79%
  - Zelluläre Rolle: 93%
- Die Bedingungen sind verletzt, wenn es eine kleine # mit Funktion ist
- Z.B. betrachten wir die funktionelle Klasse „Zelluläre Rolle“: alle Funktionsklassen erfüllen die Bedingungen, außer die Klassen: Zelladhäsion, mitochondriale Transkription und Translation
- Die # an Proteinen, die die korrespondierende Funktion haben, ist sehr klein
- geschätzte Parameter zu ungenau, werden nicht weiter betrachtet
- Alternativ kann man auch sagen, dass ein unstudiertes Protein eine bestimmte Funktion besitzt, wenn  $W$ keit eine Threshold-Funktion erfüllt

# Resultate

- Getestet wird an Interaktionspartnern, die beide studiert sind und einer von denen als unstudiert genommen wird
- Sei  $n_i$  die # an Funktionen für Protein  $P_i$  in YPD,  $m_i$  die # vorhergesagter Funktionen und  $k_i$  die Überschneidung von beobachtet/vorhergesagt

# Resultate

$$SP = \frac{\sum_i^K k_i}{\sum_i^K m_i}$$

$$SN = \frac{\sum_i^K k_i}{\sum_i^K n_i}$$

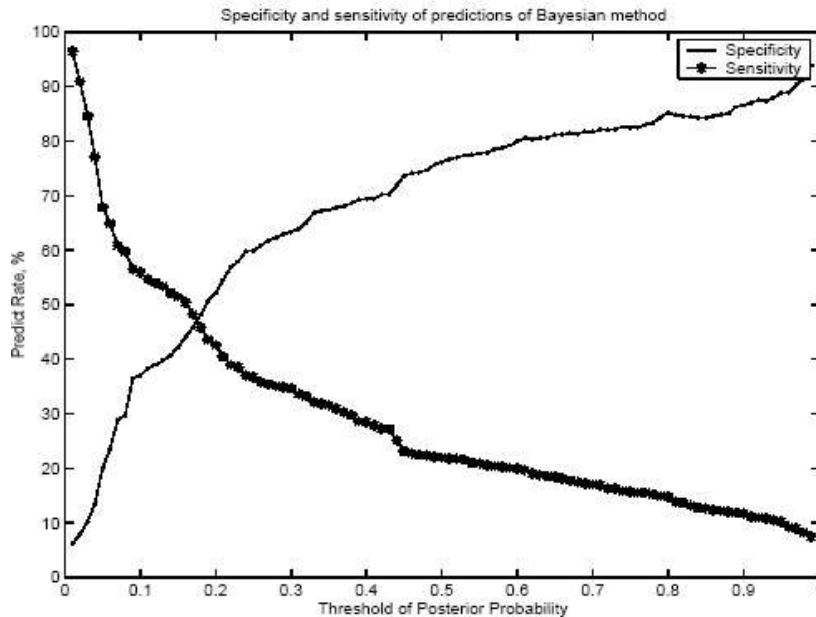


Figure 1. Specificity and sensitivity of Bayesian predictions for different thresholds.

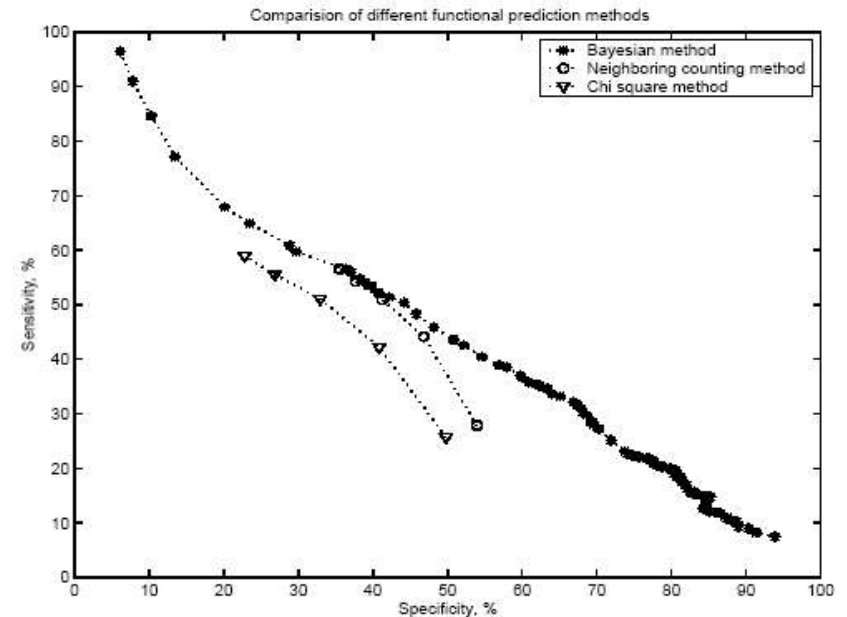
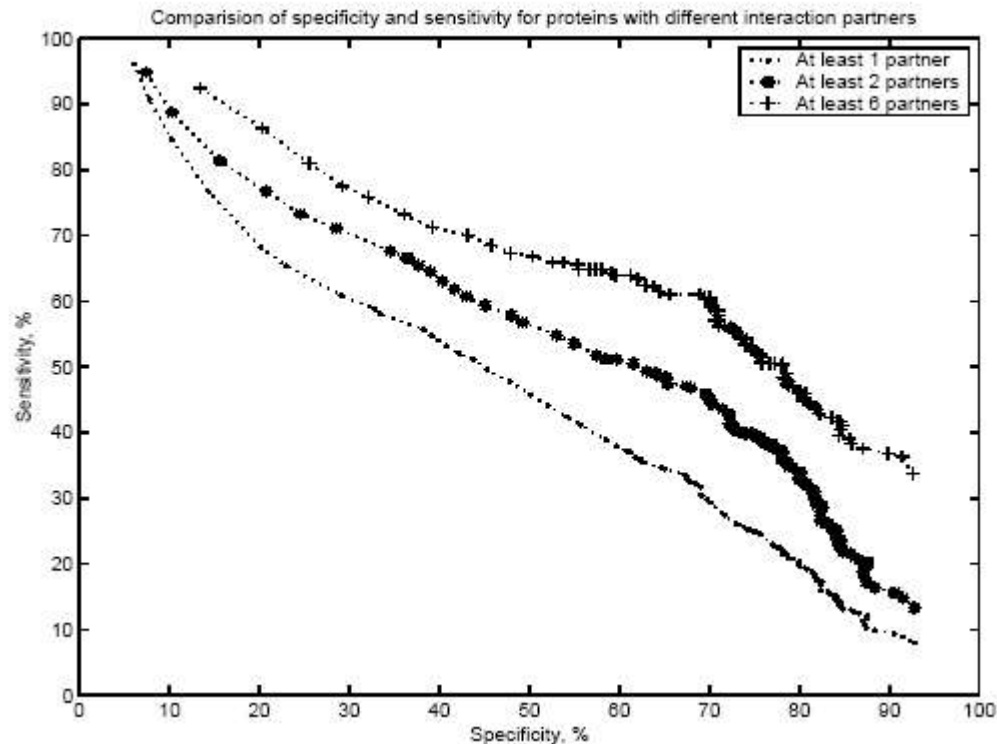


Figure 2. Sensitivity and specificity of predictions for three different methods.

# Resultate

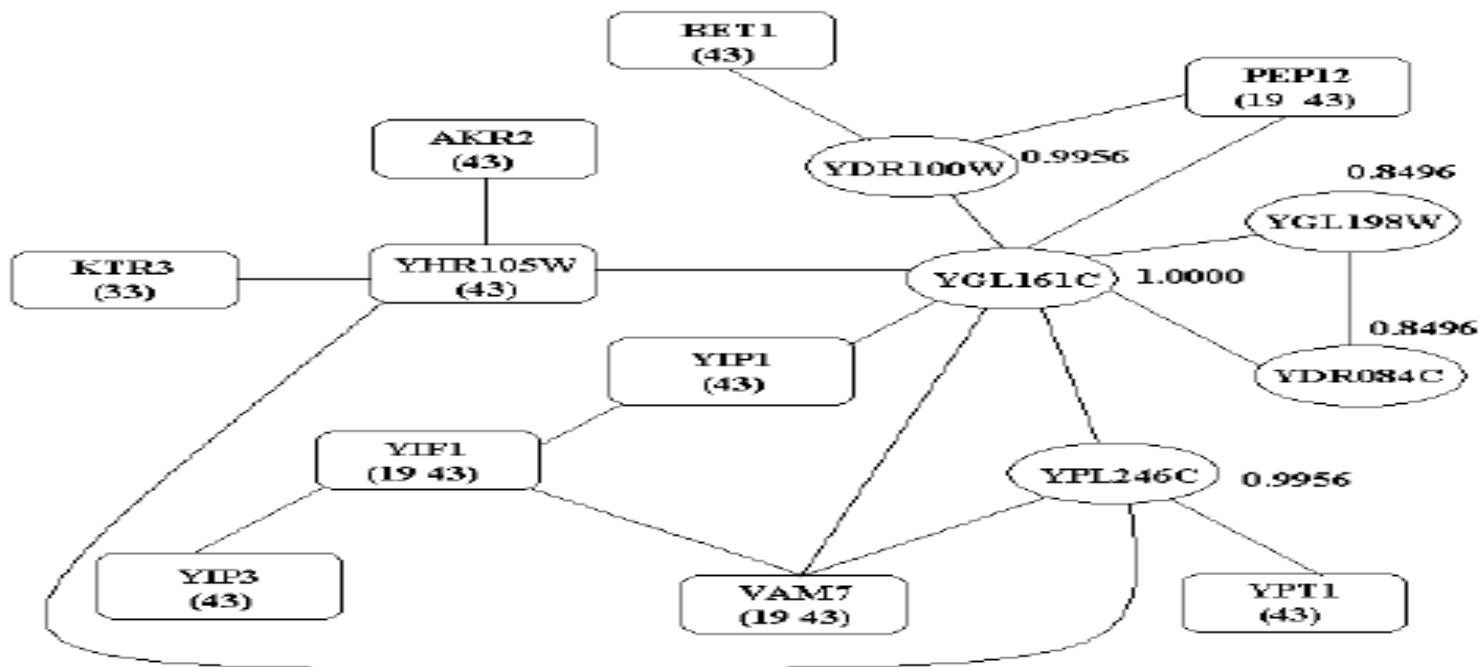
- Je mehr Interaktionspartner, desto akkurater sind die Vorhersagen, desto höher Sensitivität



# Resultate

- Es wird auch Information von indirekten Interaktionspartnern benutzt
- YDR084C hat keine studierten Partner, dafür aber YGL161C, welches 4 Partner mit Funktion 43 hat (Vesikeltransport)
- Daraus können wir zu ~84% schließen, dass YDR084C auch diese Funktion hat

# Resultate



# Literatur

- [1] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J.. 1997. Gapped BLAST and PSI-BLAST: a new generation of
- [2] Bader, G.D., Donaldson, I., Wolting, C., Ouellette, B.F., Pawson, T., Hogue, C.W.. 2001. BIND–The Biomolecular interaction network database. *Nucleic Acids Research* **29**: 242 – 245.
- [3] Brown, M., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M. Jr., and Hausler, D.. 2000. Knowledge-based analysis of microar-