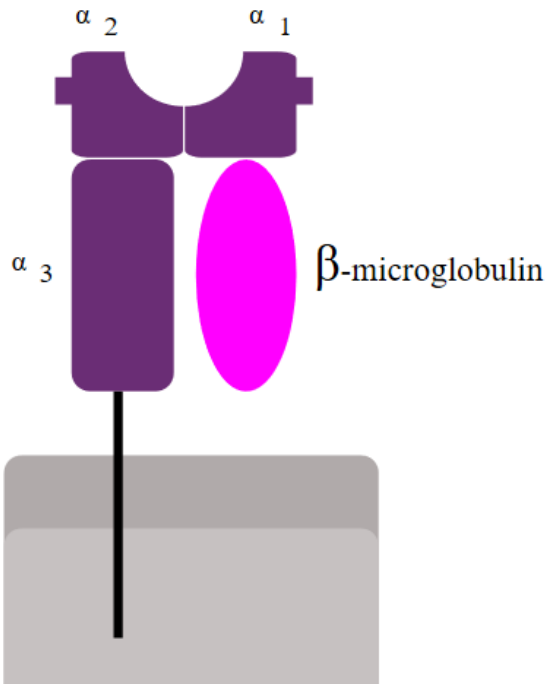
NEURONALES NETZ ZUR MHC-I BINDEVORHERSAGE FÜR EIN SPEZIFISCHES MHC ALLEL

Julian Müller, Tobias Nietsch und Lisa Falk

Computational Immunomics
04.07.2017

Aufgabenstellung



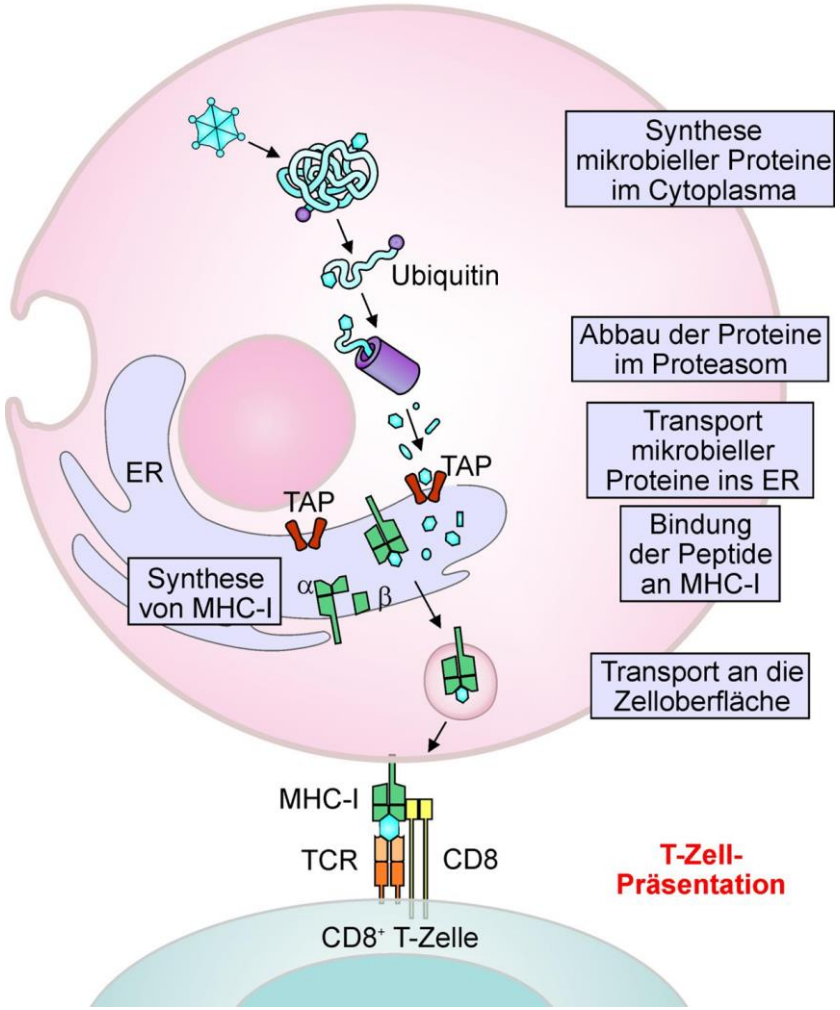
MHC-I

NFFHASLAY
YQKKNASVY
AAFLDDNAF
IQNKLSSTF
KLAEIFQPF
FAAAAARTL
MQQSGDEAF
KLRKKSSFY
GLGGDASAY
LQKVPHTRY
IQAGVDRFY
FQVNRFTGY
YFRNSGMTY
TVSPSAPTY
KTIQGGLGW
...

Bindung?

https://upload.wikimedia.org/wikipedia/commons/e/ee/MHC_Class_1.svg

MHC-Bindung


Synthese mikrobieller Proteine im Cytoplasma
Ubiquitin
Abbau der Proteine im Proteasom
Transport mikrobieller Proteine ins ER
TAP
TAP
ER
Bindung der Peptide an MHC-I
Synthese von MHC-I
α
β
Transport an die Zelloberfläche
MHC-I
TCR
CD8
T-Zell-Präsentation
CD8⁺ T-Zelle

Aufbereitung der Daten

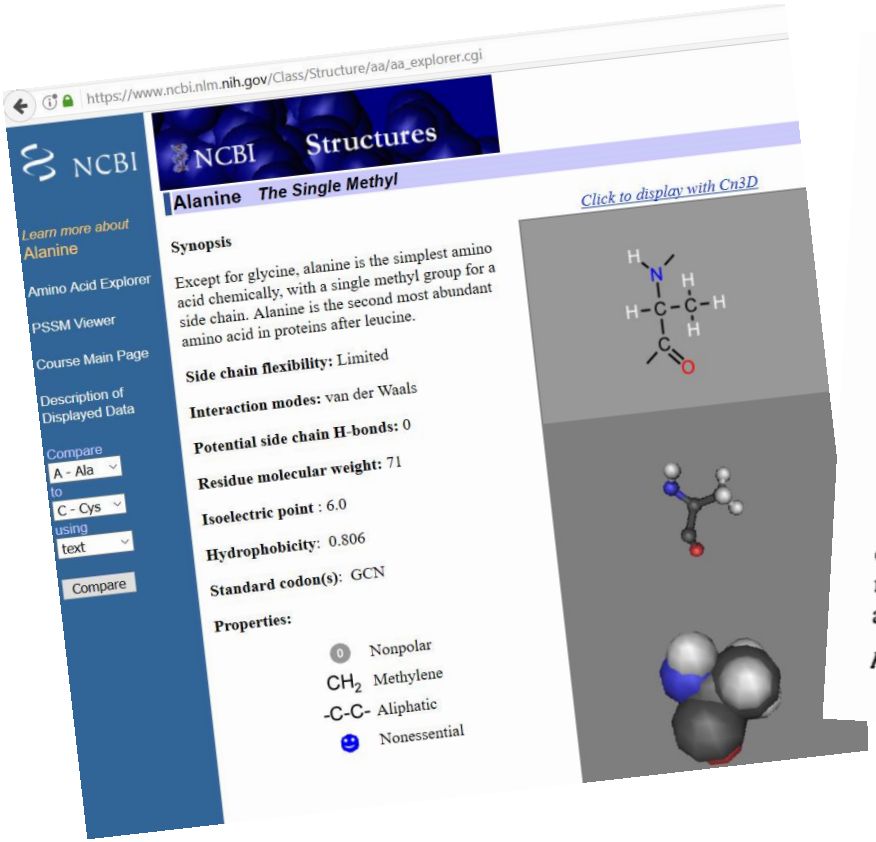
Aminosäure	Position								
	1	2	3	4	5	6	7	8	9
A	15	7	13	19	19	16	16	29	2
C	0	0	2	1	5	3	3	2	0
D	0	0	4	15	9	14	4	11	0
E	4	2	7	26	9	6	3	13	0
F	13	3	25	1	4	9	5	4	44
G	12	1	9	17	14	27	13	12	0
H	2	1	4	7	3	4	2	2	2
I	21	6	14	3	8	6	11	1	3
K	21	0	13	6	10	4	10	8	1
L	13	38	17	14	8	8	12	6	13
M	3	25	7	2	3	1	8	2	7
N	6	0	8	6	11	12	11	3	0
P	1	0	1	18	14	3	9	9	0
Q	6	54	10	8	3	5	6	8	0
R	12	0	12	7	11	11	10	6	0
S	7	10	6	4	12	18	13	17	0
T	8	11	4	4	10	12	15	17	3
V	12	17	7	6	8	6	12	17	4
W	5	0	3	4	6	3	4	3	8
Y	14	0	9	7	8	7	8	5	88

Häufigkeitsmatrix der Aminosäuren an den verschiedenen Positionen der Bindersequenzen

➡ Erkennen der Ankerpositionen an Position 2 und 9

Aufbereitung der Daten

Gliedern der AS in den Sequenzen nach Seitenkettengewicht, Isoelektrischer Punkt, Hydrophobizität, Polarität und Kontaktfläche.



Positional flexibilities of amino acid residues in globular proteins

R. BHASKARAN and P.K. PONNUSWAMY

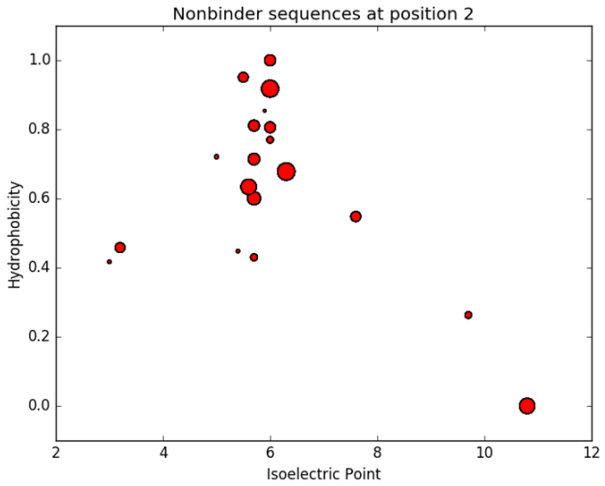
Department of Physics, Bharathidasan University, Tiruchirapalli, India

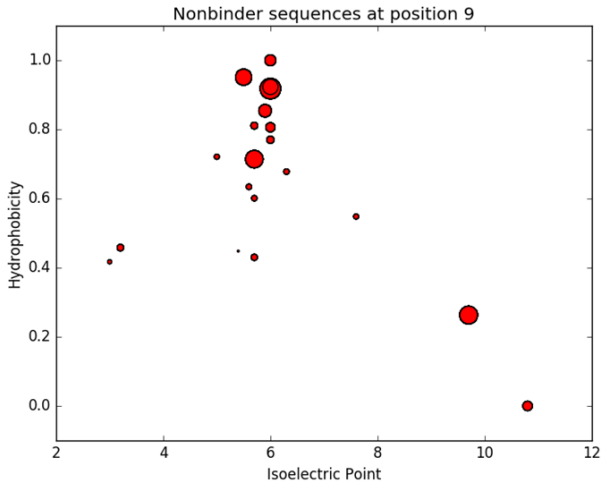
Received 7 May 1987, accepted for publication 8 March 1988

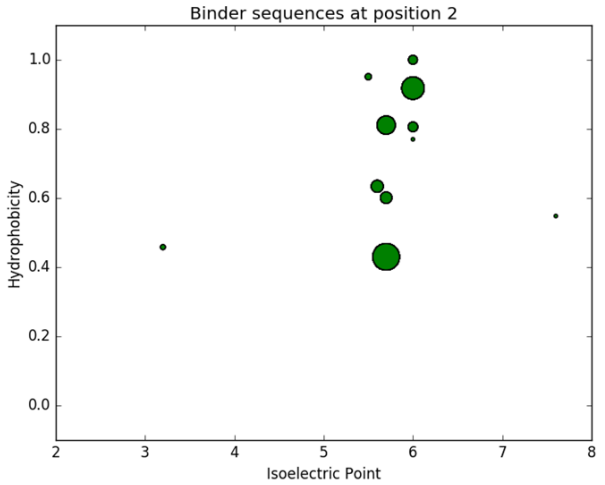
The fluctuational amplitudes of the amino acid residues in 19 protein molecules are computed by the use of our differential equation model and are analysed statistically to get new information as to their stability, position dependent nature, distribution and group dynamical behaviour, etc. The symmetric/asymmetric distribution of amino acid residues in the protein molecules is described by introducing a parameter called the "flexibility index" for each of the residue types. Finally, the overall flexible nature of the amino acid residues with respect to their spatial positioning is explained and their relationships with the residue properties are derived.

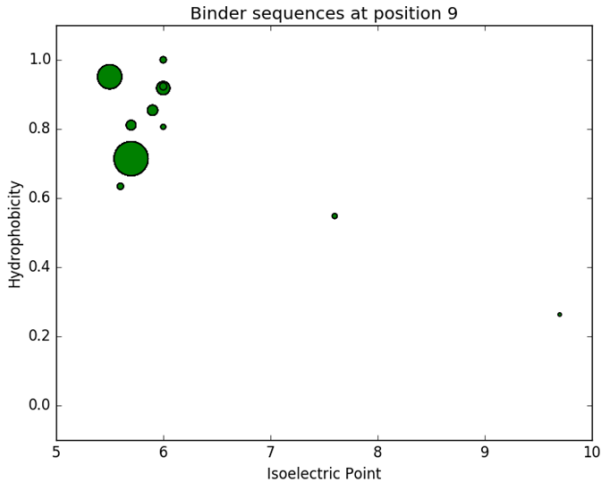
Key words: differential equation model; flexibility index; fluctuational amplitudes; spatial positions; symmetric/asymmetric distribution

Aufbereitung der Daten





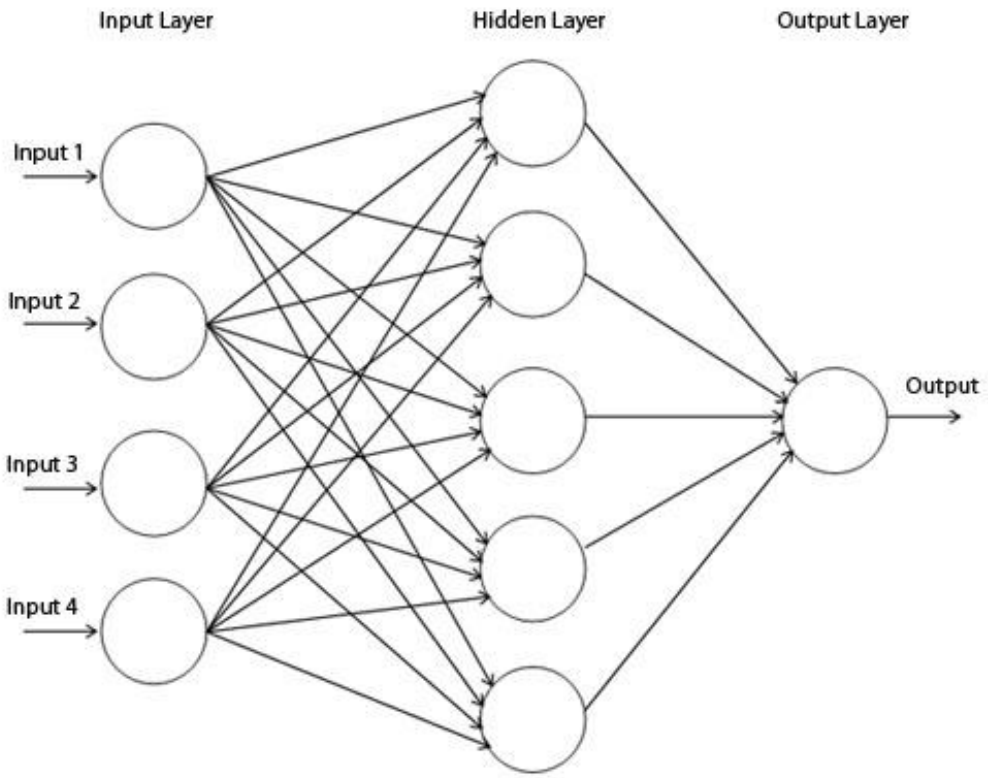




Vergleich der Streudiagramme Hydrophobizität gegen isoelektrischer Punkt für Binder und Nichtbinder an Position 2 und 9 der Aminosäuren

Konstruktion des KNN

Multilayer-Perzeptron:

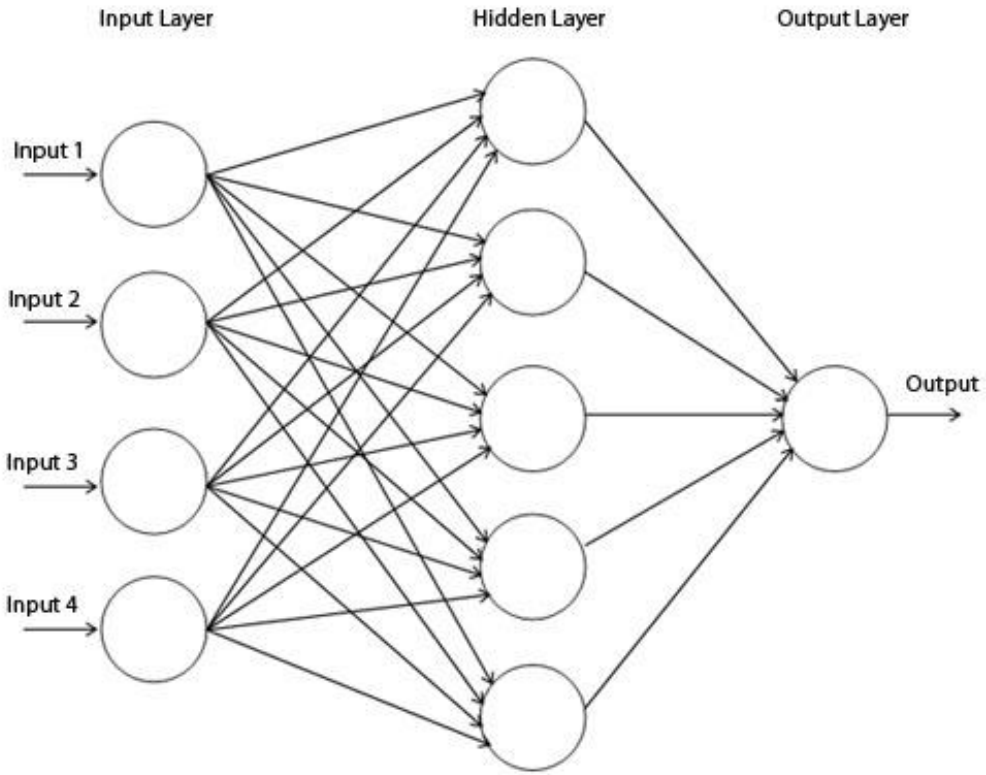


http://www.codeproject.com/KB/dotnet/predictor/network.jpg

Konstruktion des KNN

Multilayer-Perzeptron:

45 Neuronen ***1 Neuron***

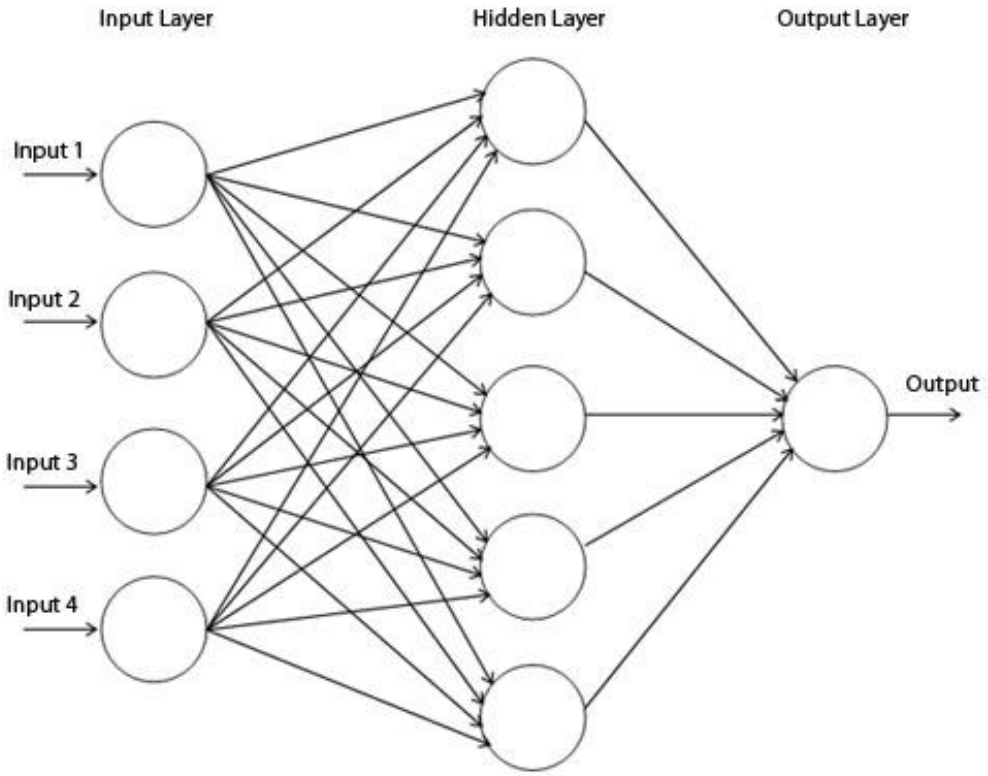


http://www.codeproject.com/KB/dotnet/predictor/network.jpg

Konstruktion des KNN

Multilayer-Perzeptron:

45 Neuronen ***??*** ***1 Neuron***



http://www.codeproject.com/KB/dotnet/predictor/network.jpg

Konstruktion des KNN

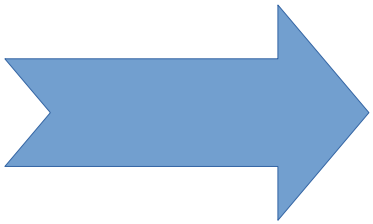
Bereits bekannte Parameter:

Parameter	*Wert*
Größe des Input Layer	45
Größe des Output Layer	1
Aktivierungsfunktion des Output Layer	Sigmoid
Zahl der Hidden Layer	1
Initialisierung der Gewichte	Uniform

Konstruktion des KNN

Unbekannte Parameter:

Parameter	*Mögliche Werte*
Größe des Hidden Layer	5 - 19
Aktivierungsfunktion von Input & Hidden Layer	8 mögliche Funktionen
Batch Size	1, 10, 100, ...
Anzahl Trainingsepochen	10, 20, 50, ...



Mehrere 1000 mögliche Parameterkombinationen!

*Daher: **2** Gridsearch-Durchläufe*

Konstruktion des KNN

1. Grid Search

- Kleiner Trainingsdatensatz (0.2 bzw. 145 Daten)
- 20 Epochen
- Ziele:
  - Wählen der Größe des Hidden Layers
  - Wählen der Aktivierungsfunktion v. Input & Hidden Layer
  - Eingrenzen der Batch Size

Konstruktion des KNN

1. Grid Search - Ergebnisse

- 17 Neuronen im Hidden Layer
- Aktivierungsfunktion ***softplus***:

$$f(x) = \ln(1 + e^x)$$

- Batch Size ≥ 100

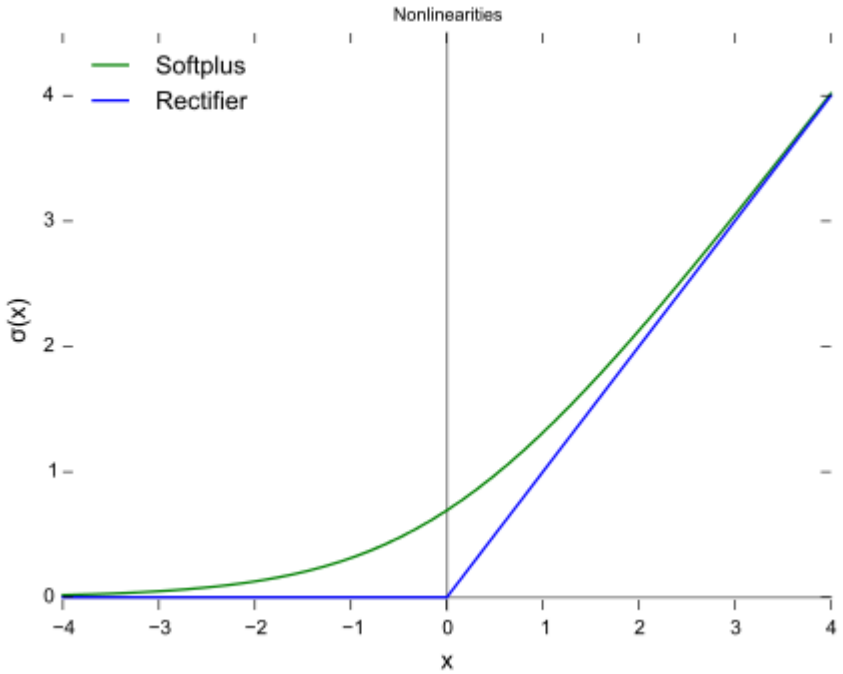
Konstruktion des KNN

1. Grid Search - Ergebnisse

- 17 Neuronen im Hidden Layer
- Aktivierungsfunktion ***softplus***:

$$f(x) = \ln(1 + e^x)$$

- Batch Size ≥ 100



Konstruktion des KNN

2. Grid Search

- Trainingsdatensatz: 0.66 bzw. 486 Trainingsdaten
- Ziele:
  - Wahl der Batch Size (100, 200, 300, 400, 486)
  - Wahl der Epochenanzahl (10, 20, 50, 100, 200, 250)

Konstruktion des KNN

2. Grid Search - Ergebnisse

		Batch Size				
	ROC_AUC	100	200	300	400	486
Epochen	10	0.818	0.805	0.745	0.717	0.655
	20	0.834	0.829	0.816	0.794	.761
	50	0.852	0.819	0.850	0.826	0.815
	100	0.902	0.860	0.854	0.817	0.801
	200	0.968	0.899	0.911	0.900	0.864
	250	0.987	0.964	0.929	0.910	0.892

Konstruktion des KNN

2. Grid Search - Ergebnisse

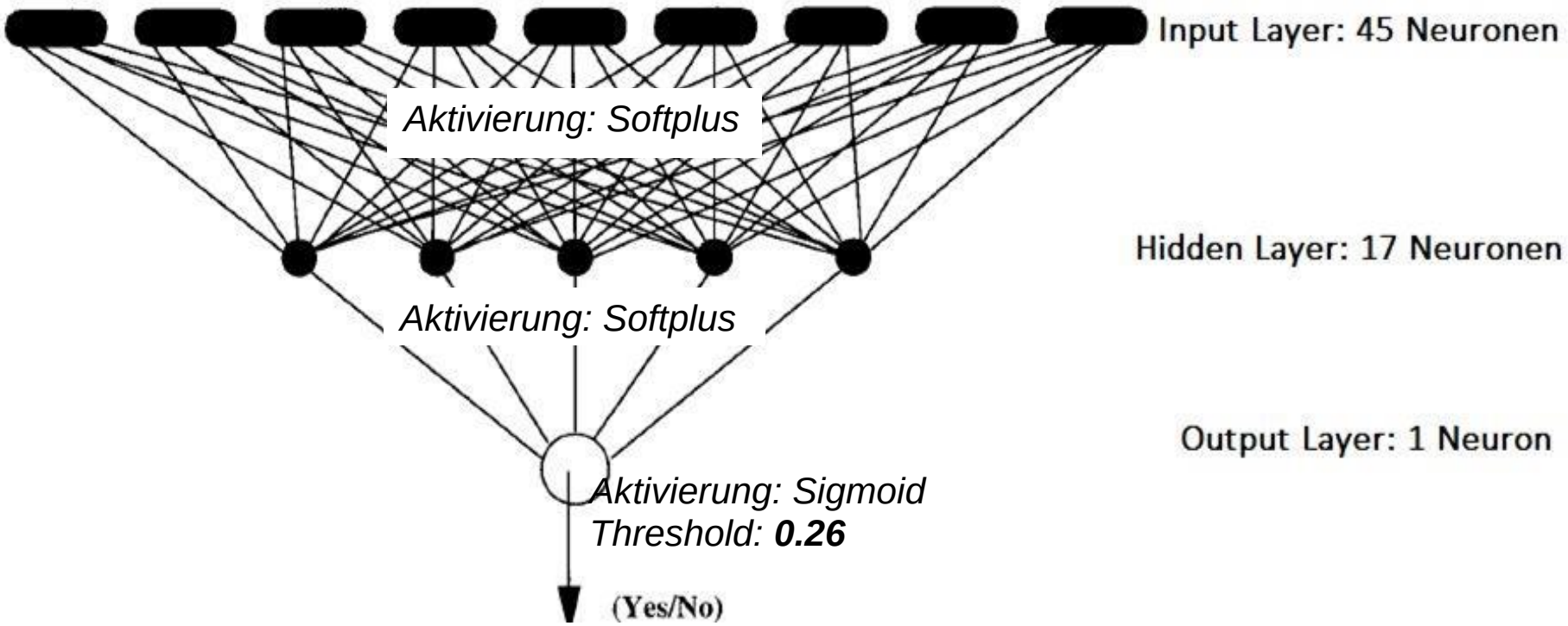
		Batch Size				
	ROC_AUC	100	200	300	400	486
Epochen	10	0.818	0.805	0.745	0.717	0.655
	20	0.834	0.829	0.816	0.794	.761
	50	0.852	0.819	0.850	0.826	0.815
	100	0.902	0.860	0.854	0.817	0.801
	200	0.968	0.899	0.911	0.900	0.864
	250	0.987	0.964	0.929	0.910	0.892

Konstruktion des KNN

2. Grid Search - Ergebnisse

		Batch Size				
	ROC_AUC	100	200	300	400	486
Epochen	10	0.818	0.805	0.745	0.717	0.655
	20	0.834	0.829	0.816	0.794	.761
	50	0.852	0.819	0.850	0.826	0.815
	100	0.902	0.860	0.854	0.817	0.801
	200	0.968	0.899	0.911	0.900	0.864
	250	0.987	0.964	0.929	0.910	0.892

Das fertige KNN



Basierend auf Gulukota et al: Two Complementary Methods for Predicting Peptides Binding Major Histocompatibility Complex Molecules J. Mol. Biol. (1997)

Evaluierung

- Konfusionsmatrix:

		Tatsächl. Wert	
		Binder	**Nichtbinder**
Vorhersage	**Binder**	36	37
	Nichtbinder	16	151

→ Genauigkeit: 77,9%

→ Sensitivität: 69,2%

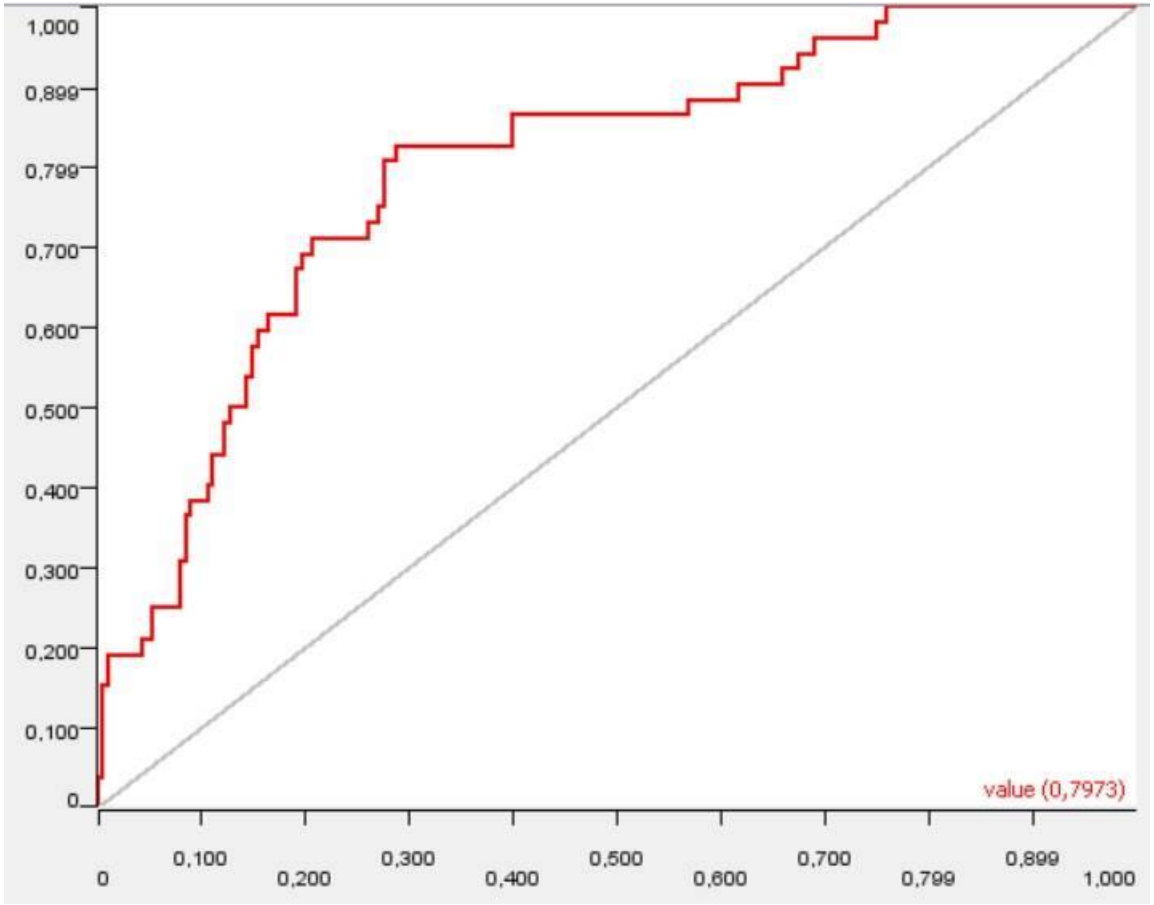
→ Spezifität: 80,3%

→ MCC: 0,443

Evaluierung

- 10-fache stratifizierte Kreuzvalidierung
- 100 Iterationen
- → Mittlere Genauigkeit von 78,93%
- → Standardabweichung von 4,14%

ROC-Kurve


1,000
0,899
0,799
0,700
0,600
0,500
0,400
0,300
0,200
0,100
0
value (0,7973)
0
0,100
0,200
0,300
0,400
0,500
0,600
0,700
0,799
0,899
1,000

Diskussion

- Ankerpositionen an Positionen 2 und 9
- → Mögliche Verbesserung durch Einbeziehen
- Paper von Brusic et al. liefert nur leicht bessere Werte
- Bei anders verteiltem Datensatz möglicherweise andere Performance
- AUC von 0,79 schließt Zufallsprediktor aus