EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Einführung in die Bioinformatik:

Lernen mit Kernen

Prof. Dr. Karsten Borgwardt

Data Mining in den Lebenswissenschaften

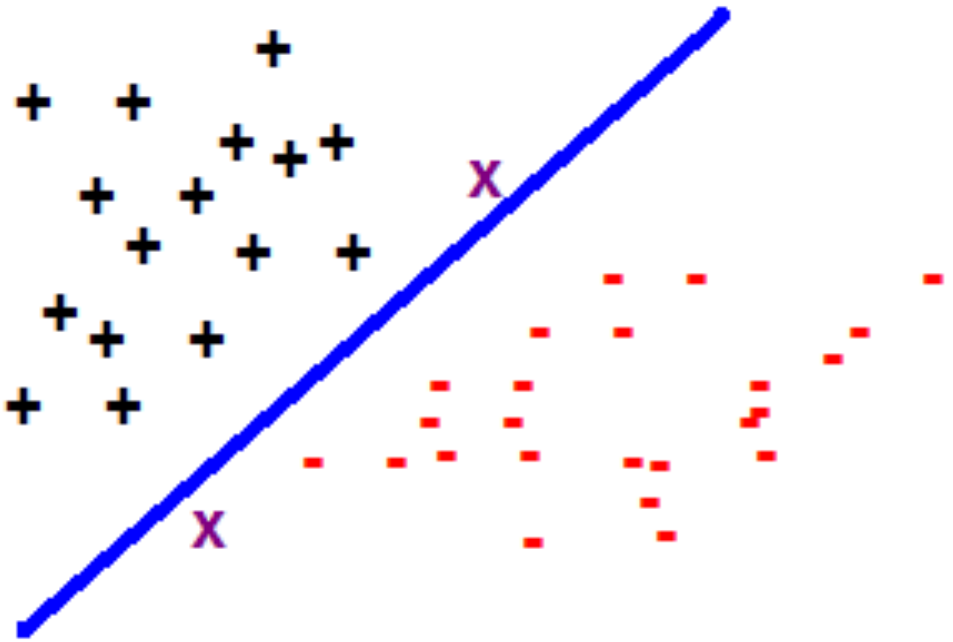
Universität Tübingen &

Max-Planck-Institut für Intelligente Systeme &

Max-Planck-Institut für Entwicklungsbiologie

Morgenstelle N1, 11.7.2013

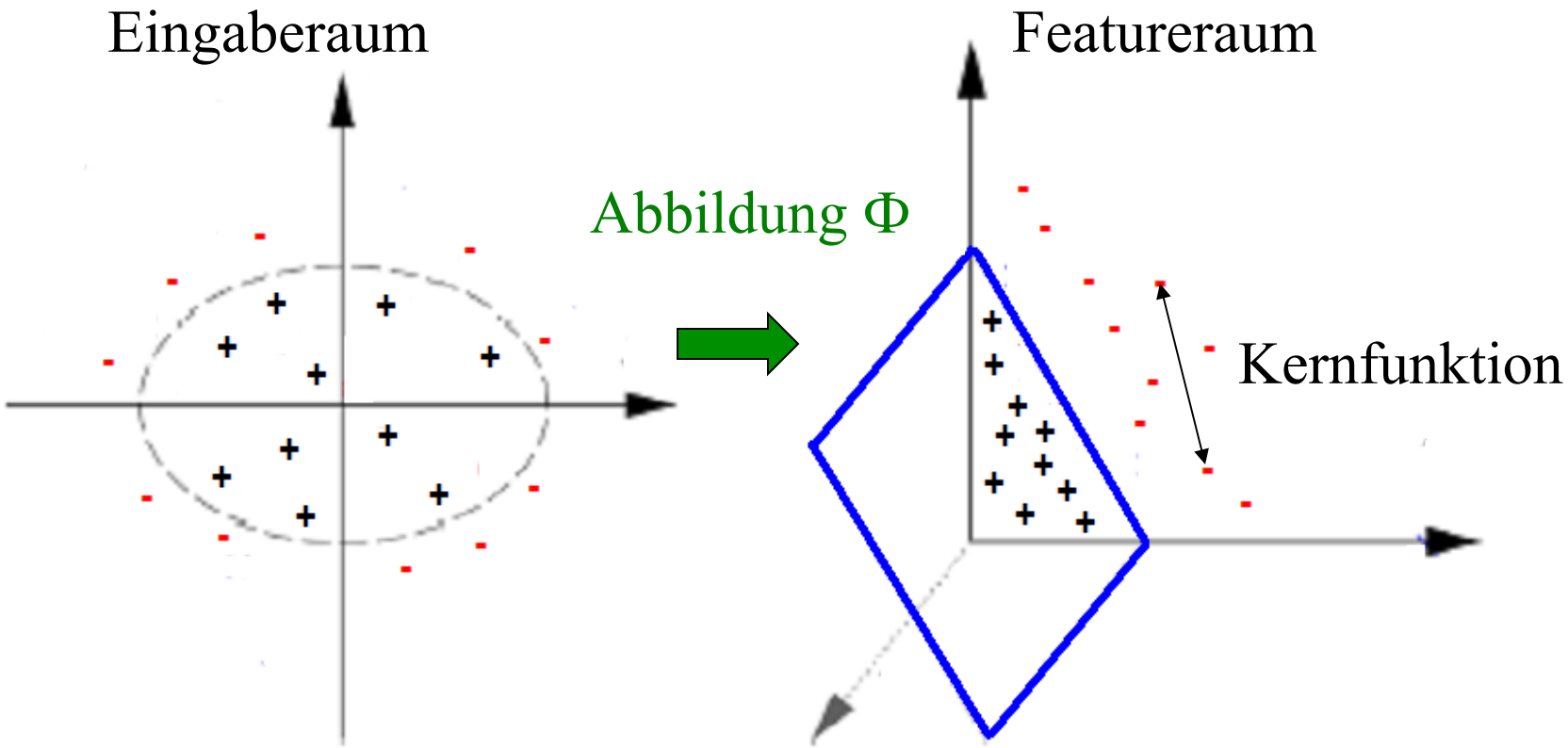
Support Vector Machines



Sind neue Datenpunkte (**x**) rot oder schwarz?

Die blaue Hyperebene dient der Vorhersage der Klassenzugehörigkeit neuer Punkte.

Kern-Trick



Der Kern-Trick erlaubt die effiziente Berechnung einer trennenden Hyperebene im Featureraum.

Klassifikation

Wie funktioniert die SVM-Klassifikation?

Wir berechnen die Entscheidungsfunktion f:

$$f(x) = sgn(\langle w, \phi(x) \rangle + b)$$

wobei x ein Datenpunkt, w der Gewichtsvektor der Hyperebene und b eine Konstante ist.

Klassifikation

Seien $\{x_1,...,x_n\}$ die Trainingspunkte, $\{y_1,...,y_n\}$ ihre Klassenlabels (+1 oder -1).

Dann ist die Klassifikationsregel äquivalent zu:

$$f(x) = sgn(\sum_{i=1}^{n} \alpha_i y_i < \phi(x_i), \phi(x) > +b) =$$

$$= sgn(\sum_{i=1}^{n} \alpha_i y_i k(x_i, x) + b)$$

wobei $k(x, x_i) =< \phi(x_i), \phi(x) >$

die sogenannte **Kernfunktion (der „Kern“)** ist.

Kerne

Linearer Kern: $k(x, x') =< x, x' >= \sum_{j=1}^{d} x^{(j)} * x'^{(j)}$

Polynomieller Kern: $k(x, x') = (< x, x' > +c_1)^{c_2}$

Gauß-Kern: $k(x, x') = \exp(-\gamma||x - x'||^2)$

Delta-Kern: $k(x, x') = \begin{cases} 1, & \text{wenn x = x'} \\ 0, & \text{wenn x} \neq \text{x'} \end{cases}$

c_1, c_2 und γ sind positive Skalare.

Abgeschlossenheit von Kernen

Kerne sind abgeschlossen unter Addition und punktweiser Multiplikation:

Additivität:

Falls k ein Kern und l ein Kern ist, dann ist auch k+l ein Kern.

Multiplikativität:

Falls k ein Kern ist und l ein Kern ist, dann ist auch k*l ein Kern.

Karsten Borgwardt

Protein function prediction via graph kernels

in Zusammenarbeit mit

Cheng Soon Ong and S.V.N. Vishwanathan,

Stefan Schönauer, Hans-Peter Kriegel und Alex Smola

ISMB 2005

Inhalt

Einführung

- **Das Problem: Proteinfunktionsvorhersage**
- **Die Methode: Support Vector Machines (SVM)**

Unser Ansatz zur Funktionsvorhersage

- **Graphenmodell für Proteine**
- **Graphkern für Proteine**
- **Experimentelle Ergebnisse**

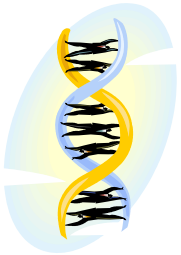
Techniken zur Vorhersageverbesserung

- **Hyperkerne**

Diskussion

Proteinfunktionsvorhersage

Molekularer Informationsfluss



Gen

Vorhersage

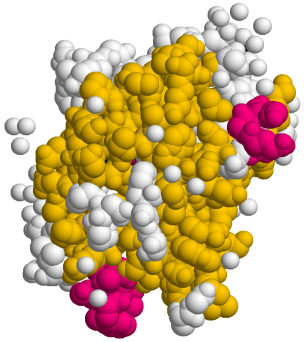
Proteinsequenz

SSFSWDNCDEGKDPAVIRSLTLEPDPI
IVPGNVTLSVMGSTSVPLSSPLKVDL
VLEKEVAGLWIKIPCTDYIGSCTFEHF
CDVLDMLIPTGEPCPEPLRTYGLPCH
CPFKEGTYSLPKSEFVVPDLELPSWL
TTGNYRIESVLSSSGKRLGCIKIAASL
KGI

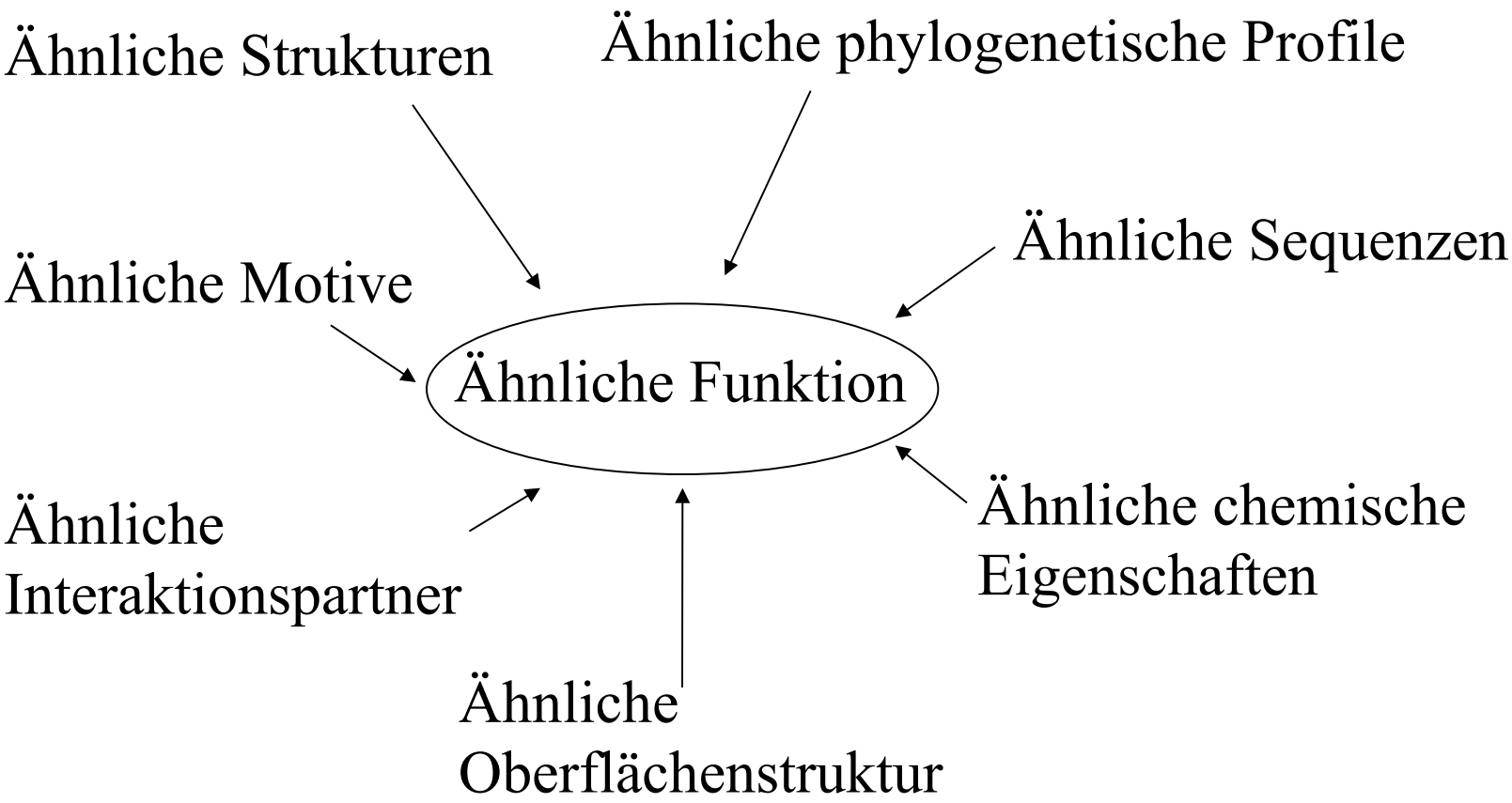
Proteinfunktion

z.B. Enzym

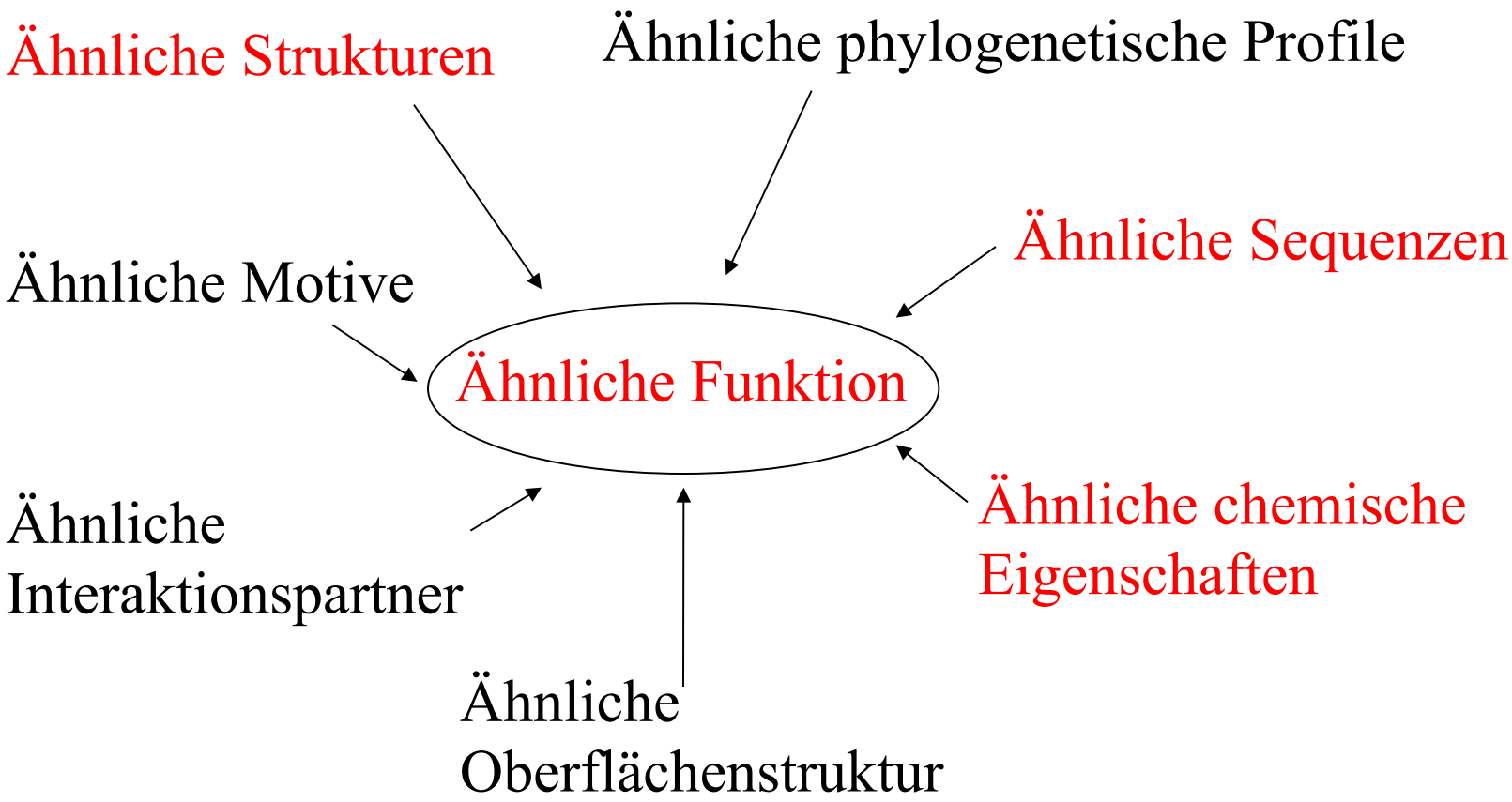
Proteinstruktur



Bekannte Ansätze zur Proteinfunktionsvorhersage

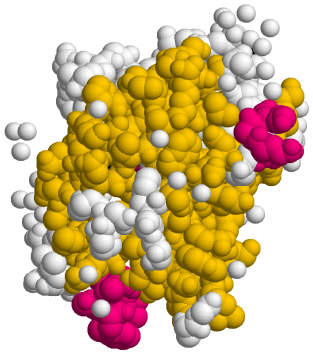


Bekannte Ansätze zur Proteinfunktionsvorhersage



Featurevektoren zur Funktionsvorhersage

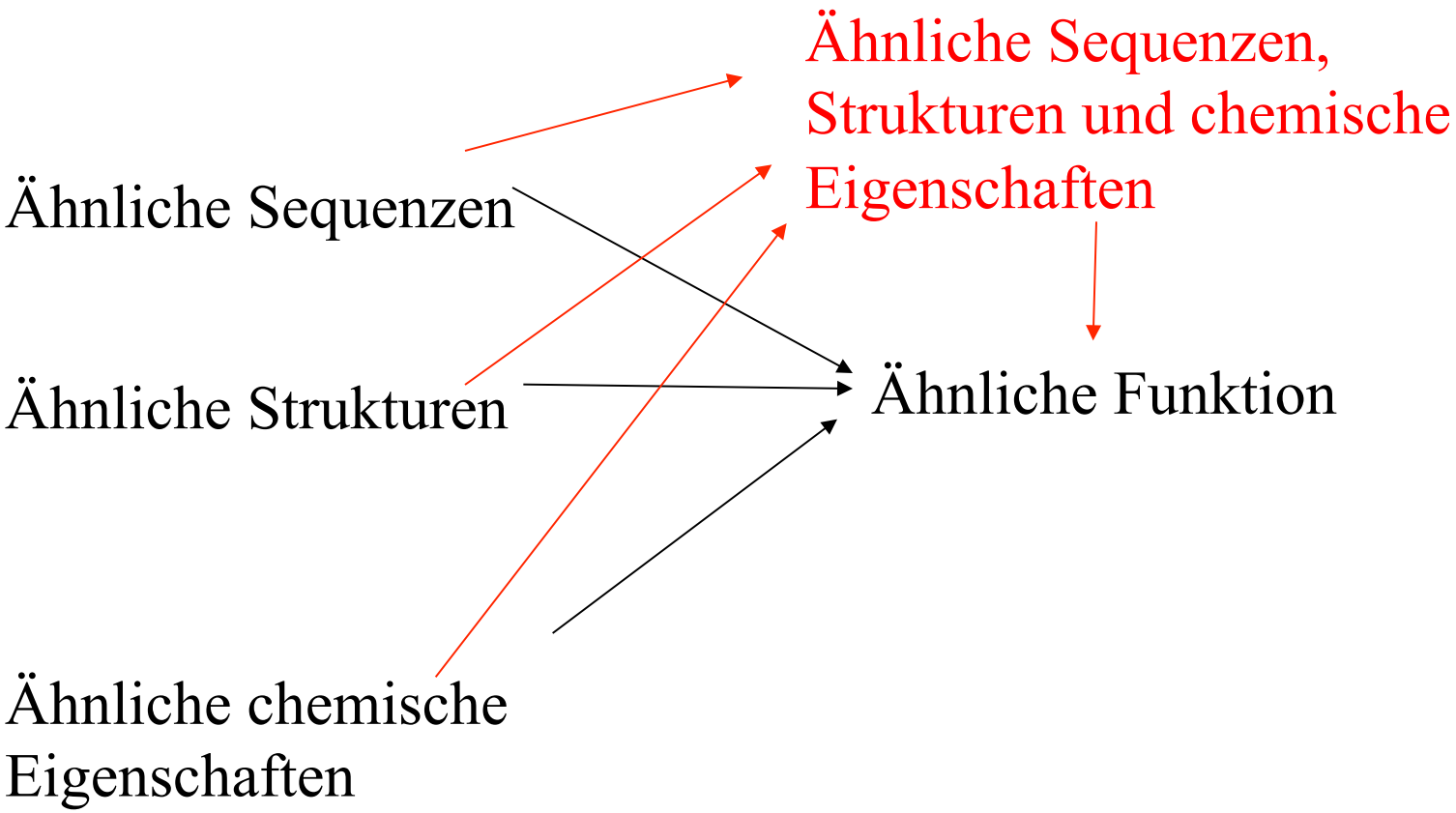
Proteinstruktur
und/oder
Proteinsequenz



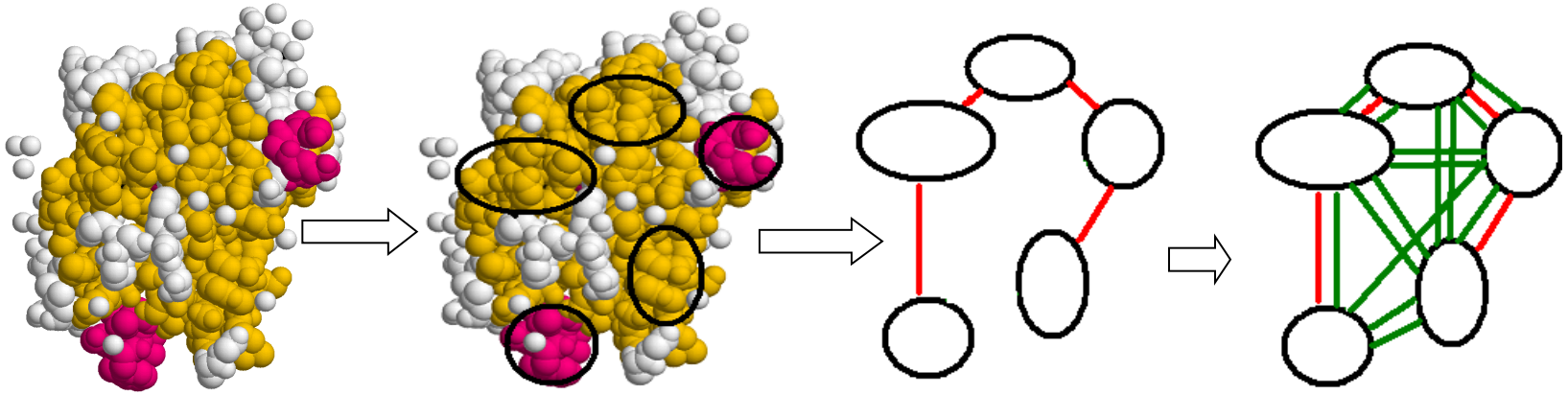
z.B. Dobson and Doig (2003),
Cai et al. (2004)

- Hydrophobizität
- Polarität
- Polarisierbarkeit
- Van-der-Waals-Volumen
- Histogramm über Aminosäuretypen
- Histogramm über Oberflächenanteile
- Disulfid-Bindungen

Unser Ansatz



Graphenmodell für Proteine



Protein

Sekundär-
struktur

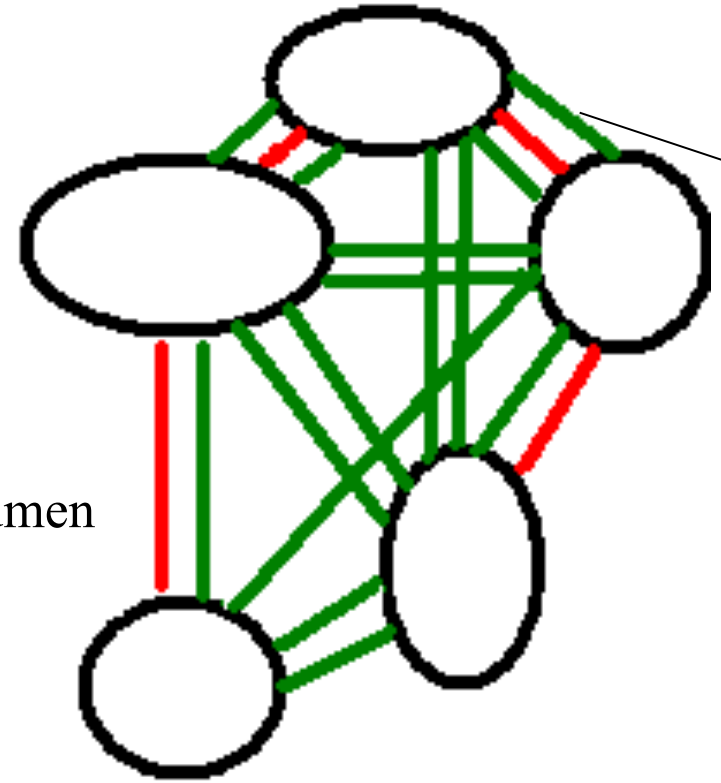
Sequenz

3D-Struktur

Graphenmodell für Proteine

Knotenattribute

- Hydrophobizität
- Polarität
- Polarisierbarkeit
- Van-der-Waals-Volumen
- Länge
- Sekundärstruktur



Kantenattribute

- Typ (Sequenz, Struktur)
- Länge

Kerne auf Wegen

vergleichen Wege identischer Länge und erweitern die Kerne von Kashima et al. (2003) und Gärtner et al. (2003)

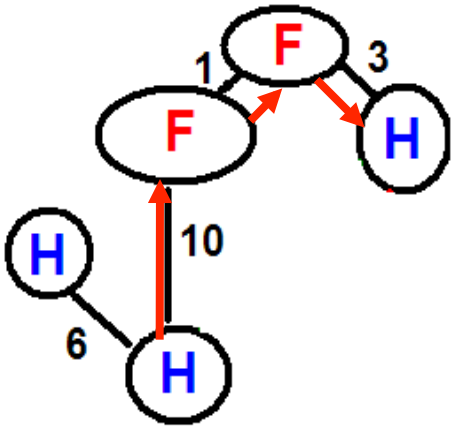
$$k_{walk}((v_1,\ldots,v_l),(w_1,\ldots,w_l)) = \prod_{i=1}^{l-1} k_{step}((v_i,v_{i+1}),(w_i,w_{i+1}))$$

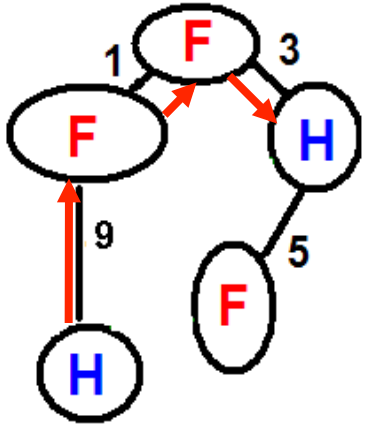
Zwei Wege sind ähnlich, wenn entlang dieser Wege

- **die Typen von Sekundärstrukturelementen** (SSE) identisch sind,
- **die Distanzen** zwischen SSE ähnlich sind,
- **die chemischen Eigenschaften** von SSE ähnlich sind.

Kerne auf Wegen

Protein A



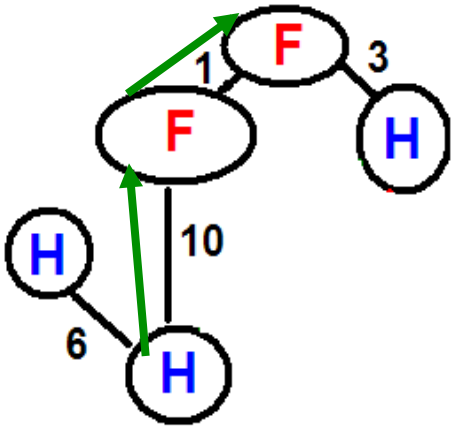


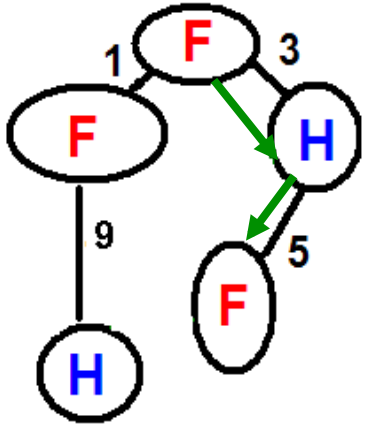
Protein B

Ähnlich (H,10,F,1,F,3,H) (H,9,F,1,F,3,H)

Kerne auf Wegen

Protein A





Protein B

Unähnlich (H,10,F,1,F) (F,3,H,5,F)

Evaluation: Enzyme versus Nicht-Enzyme

10-fach Kreuzvalidierung auf 1128 Proteinen (Dobson and Doig, 2003); 59 % sind Enzyme.

Kernel	Genauigkeit	Standardabweichung	
Vector kernel	76,86	1,23	
Optimized vector kernel	80,17	1,24	
Graph kernel	77,30	1,20	
Graph kernel without structure	72,33	5,32	
Graph kernel with global info	84,04	3,33	
DALI classifier	75,07	4,58	

Hyperkerne

Welches Attribut ist am wichtigsten für die korrekte Klassifikation?

Um diese Frage zu untersuchen, verwenden wir Hyperkerne(Ong et. al, 2003).

Hyperkerne finden eine optimale Linearkombination einer Menge von gegebenen Kernmatrizen:

$$\sum_{i=1}^{n} \lambda_i K_i$$

Minimiert den Trainingsfehler und erfüllt Regularisierungsbedingungen.

Hyperkerne

Unser Ansatz:

- Berechne eine Kernmatrix für 600 Proteingraphen mit **nur einem Knotenattribut**.
- Wiederhole dies für alle Attribute.
- Normalisiere diese Kernmatrizen.
- Bestimme eine Hyperkern-Linearkombination.
- λ_i stellt dann den Beitrag von Feature i zur korrekten Klassifikation dar.

Hyperkerne

Attribute	EC 1	EC 2	EC 3	EC 4	EC 5	EC 6
Amino acid length	**1.00**	**0.31**	**1.00**	**1.00**	**0.73**	0.00
3-bin van der Waals	0.00	0.00	0.00	0.00	0.00	0.00
3-bin Hydrophobicity	0.00	0.00	0.00	0.00	0.00	0.00
3-bin Polarity	0.00	0.01	0.00	0.00	0.00	**1.00**
3-bin Polarizability	0.00	0.00	0.00	0.00	**0.12**	0.00
3d length	0.00	**0.40**	0.00	0.00	0.00	0.00
Total van der Waals	0.00	0.00	0.00	0.00	0.00	0.00
Total Hydrophobicity	0.00	**0.13**	0.00	0.00	0.01	0.00
Total Polarity	0.00	**0.14**	0.00	0.00	0.01	0.00
Total Polarizability	0.00	0.01	0.00	0.00	**0.13**	0.00

Diskussion

- **Neuer, kombinierter Ansatz zur Proteinfunktionsvorhersage basierend auf Sequenz, Struktur und chemischen Eigenschaften**
- **Erreicht basierend auf weniger Informationen bereits Klassifikationsergebnisse, die dem Stand der Technik entsprechen; mit der identischen Menge an Informationen erzielt er sogar höhere Genauigkeitslevel.**
- **Hyperkerne zur Suche nach den interessantesten Proteineigenschaften (und eine Methode zum Kombinieren von Kernen unter gemeinsamen Regularisierungsbedingungen [ESANN 2005])**

Diskussion

- **Detaillierte Graphmodelle (Aminosäuren, Atome) sind interessanter, führen jedoch zu Berechnungsproblemen, da die Graphen zu groß werden.**

Zwei mögliche Richtungen für zukünftige Projekte

- **Effiziente und zugleich expressive Graphkerne**
- **Integration weiterer Informationen in unser Graphenmodell**

Update: Seit 2009 können wir Kerne auch auf sehr großen Graphen (Tausende von Knoten) berechnen.

Literatur

- **Borgwardt, Ong, Schönauer, Vishwanathan, Smola, Kriegel. *Protein function prediction via graph kernels*. ISMB 2005 and Bioinformatics 2005, 21(suppl_1):i47-i56**

- **Borgwardt, *Kernel Methods in Bioinformatics*, Springer Handbooks of Computational Statistics: Statistical Bioinformatics pp 317-334**

- **Smola und Schölkopf, *Learning with Kernels*, MIT Press 2002**