Discovering structural motifs using a structural alphabet

Application to magnesium-binding sites

Sebastian Bartschat

21. Juli 2007

Inhalt

Wozu dient Magnesium?

Magnesium ist einer der vielseitigsten metallischen Co-Faktoren:

- an Reaktionen mit Phosphatgruppenübertragung beteiligt (ATP-Magnesium liegt vor)
- nötig für die Nukleinsäurebiosynthese, aber auch für die Stabilität
- Stabilisierung von Proteinstrukturen
- besitzt calciumantagonistische Wirkung

Bindungseigenschaften

- Bindung an Seitenketten von Asp und Glu bzw. Asn und Gln
- bis jetzt nur Studien mit Proteinen mit hoher Sequenzähnlichkeit
  2 Sequenzmotive :

1	NA**D**F**D**G**D**	RNA Pol. und DNA Pol. I
2	YX**DD** / LX**DD**	Reverse Transkriptase und Telomerase

Fragestellungen

1. Besitzen Mg^{2+}-Bindungsstellen strukturelle Neigungen?
2. Existieren strukturelle Motive, auch wenn keine Sequenzähnlichkeit vorhanden ist?
3. Können strukturelle Motive bestimmten Proteinfunktionen zugeordnet werden?
4. Welche Spezifität weisen die gefundenen Motive auf?

- **Ziel** : Generierung eines Sets von Strukturblöcken, um die Struktur eines Proteins anhand der Sequenz bestmöglich zu approximieren

- **Ziel** : Generierung eines Sets von Strukturblöcken, um die Struktur eines Proteins anhand der Sequenz bestmöglich zu approximieren
- **Vorüberlegung** : Zerlegung des Proteins in überlappende Blöcke von M = 5 AS (ProteinBlock - PB)

- **Ziel** : Generierung eines Sets von Strukturblöcken, um die Struktur eines Proteins anhand der Sequenz bestmöglich zu approximieren
- **Vorüberlegung** : Zerlegung des Proteins in überlappende Blöcke von M = 5 AS (ProteinBlock - PB)
- **ProteinBlock** : Beschreibung durch Vektor aus Diederwinkeln

$$V(\psi_{n-2}, \phi_{n-1}, \psi_{n-1}, \phi_n, \psi_n, \phi_{n+1}, \psi_{n+1}, \phi_{n+2})$$

um das zentrale $C\alpha_n$ des ProteinBlocks

- self-organizing maps (SOM) mit 2 Lernphasen
- Ähnlichkeitsmessung zwischen 2 Vektoren mittels RMSDA (root mean square deviation on angular values)

$$\mathrm{RMSDA}(V_1, V_2) = \sqrt{\frac{\sum_{i=1}^{i=M-1} [\psi_i(V_1) - \psi_i(V_2)]^2 + [\phi_{i+1}(V_1) - \phi_{i+1}(V_2)]^2}{2(M-1)}}$$

- ProteinBlöcke PB_k der Karte sind zu Beginn durch einen Vektor $W(k)$ definiert

first training

- Einlesen von C Proteinen als konsekutive Signale der einzelnen PBs
- Bestimmung des ProteinBlocks W_k mit dem kleinsten Abstand zu dem induzierten Vektor $V(m)$ (mittels RMSDA)
- Neugewichtung des PB_k :

$$W(k) \leftarrow W(k) + (V(m) - W(k)) * \nu(c)$$

wobei $\nu(c)$ eine Lernrate darstellt, die mit der Zeit abnimmt

second training

- Ermittlung der Transitionsmatrix der PBs
- erneutes Einlesen von C Proteinen als Signal konsekutiver Vektoren
- Bestimmung von n ähnlichen Blöcken zu einem induzierten Vektor $V(m)$ (ergibt Menge von Vektoren W_k)
- Neugewichtung desjenigen $W(k)$ mit der höchsten Übergangswahrscheinlichkeit von $W(j)$
  ($W(j)$ mit $V(m-1)$ assoziiert)

shrinking process

Ziel : Reduzierung der Neurone auf eine optimale Anzahl

1. untersuche alle PB_i und PB_j der Karte K auf Ähnlichkeit $(i, j \in \{1 \ldots |K|\})$
   $\longrightarrow$ strukturelle Ähnlichkeit
   $\longrightarrow$ Translationsähnlichkeit
2. falls PB_i und PB_j ähnlich: lösche den PB der seltener ist und gehe zu Schritt 1
   sonst: ENDE

Ziel: Genaue Approximation der 3D-Struktur aus Proteinsequenz

Vorüberlegungen

- jedem ProteinBlock ist eine Menge von Sequenzen zugeordnet
- Erstellen eines Sequenzfenster der Größe $[-\omega, +\omega]$ auf der Proteinsequenz um das $C\alpha$ des jeweiligen PB

Ziel: Genaue Approximation der 3D-Struktur aus Proteinsequenz

Vorüberlegungen

- jedem ProteinBlock ist eine Menge von Sequenzen zugeordnet
- Erstellen eines Sequenzfenster der Größe $[-\omega, +\omega]$ auf der Proteinsequenz um das $C\alpha$ des jeweiligen PB
- Vorkommensmatrix für AS an Pos. j im ProteinBlock k : $n^k_{i,j}$
- Berechnung von $\mathrm{P}(a_i \text{ in } j | \mathrm{PB}_k) = \frac{n^k_{i,j}}{N_k}$

  k ... Nummer des ProteinBlocks
  i ... Nummer der Aminosäure
  j ... Position im Sequenzfenster

Beschreibung der Daten:

- Kullback-Leibler-Divergenz

$$K_k(p_j, q) = \sum_i p_{ji} * \ln\left(\frac{p_{ji}}{q_i}\right)$$

Beschreibung der Daten:

- Kullback-Leibler-Divergenz

$$K_k(p_j, q) = \sum_i p_{ji} * \ln\left(\frac{p_{ji}}{q_i}\right)$$

Unterschied zwischen der AS-Verteilung an der Stelle j im Block k zur allgemeinen Verteilung

dient der Suche nach Positionen mit großer Spezifität

- Z-Score

$$z = \frac{\left(n_{ij}^{k} - n_{ib}\right)}{\sqrt{n_{ib}}}$$

$n_{ib} = N_k * f_i$ beschreibt den Erwartungswert der i-ten AS

dient der Suche nach AS mit großer Spezifität für eine bestimmte Pos. in einem bestimmten ProteinBlock

- gesucht ist ProteinBlock k so dass $P(PB_k \mid X_s)$ maximal ist
- $P(PB_k \mid X_s) = \frac{P(X_s \mid PB_k) * P(PB_k)}{P(X_s)}$

$$P(X_s) = \prod_{j=-\omega}^{j=+\omega} P(a_i)$$

Wahrscheinlichkeit die Sequenz X_s zu beobachten; ohne gegebene Strukturinformationen

$$P(X_s) = \prod_{j=-\omega}^{j=+\omega} P(a_i)$$

Wahrscheinlichkeit die Sequenz X_s zu beobachten; ohne gegebene Strukturinformationen

$$P(X_s \mid PB_k) = \prod_{j=-\omega}^{j=+\omega} P(a_j \mid PB_k)$$

Wahrscheinlichkeit die Sequenz $X_s(a_{-\omega}, \ldots, a_{+\omega})$ zu beobachten, wenn ein bestimmter ProteinBlock k gegeben ist

Berechnung des optimalen ProteinBlockes PB* für eine gegebene Sequenz X_s mittels des Verhältnisse R_k:

$$R_k = \frac{\mathrm{P}(\mathrm{X}_s \mid \mathrm{PB}_k)}{\mathrm{P}(X_s)} = \frac{\mathrm{P}(\mathrm{PB}_k \mid X_s)}{\mathrm{P}(\mathrm{PB}_k)}$$

Berechnung des optimalen ProteinBlockes PB* für eine gegebene Sequenz X_s mittels des Verhältnisse R_k:

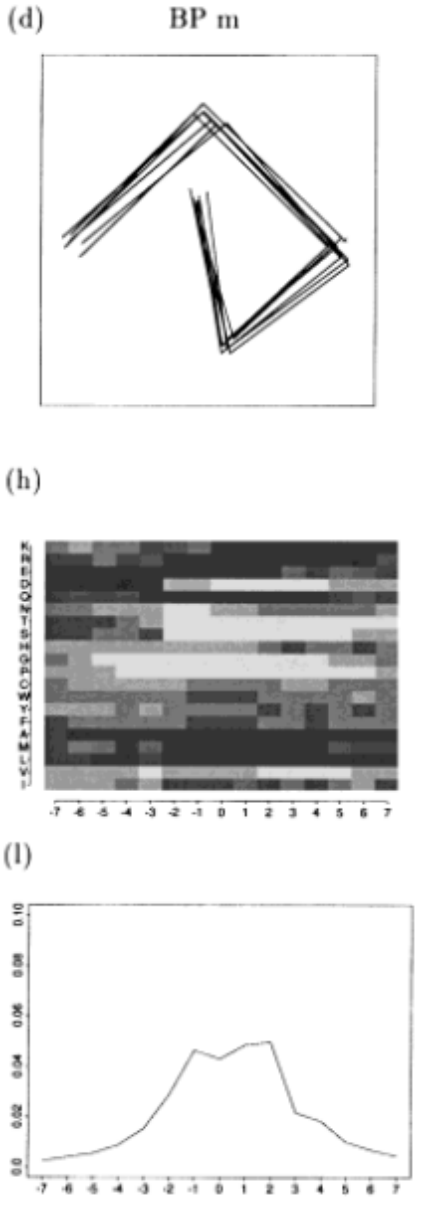
$$R_k = \frac{P(X_s \mid PB_k)}{P(X_s)} = \frac{P(PB_k \mid X_s)}{P(PB_k)}$$

1. PB* = k mit $\ln(R_k)$ maximal
2. PB* ist in der Menge der r-besten Blöcke k

- das Training resultierte in einem Alphabet aus 16 Protein-Blöcken

				Transitions (%)			Str. II (%)			
PB label	Frequency (%)	RMSD (Å)	*anr*	1st	2nd	3rd	α	Coil	β	Coarse char.
a	3.93	0.52	1.01	54.8(*c*)	16.5(*f*)	8.0(*b*)	0.1	76.7	23.3	N-cap β
b	4.58	0.51	1.00	44.4(*d*)	17.9(*c*)	13.7(*f*)	0.2	86.7	13.1	N-cap β
c	8.63	0.51	1.28	62.2(*d*)	24.4(*f*)	5.6(*e*)	0.1	58.2	41.7	N-cap β
d	18.84	0.48	2.74	51.9(*f*)	25.6(*c*)	19.2(*e*)	0.0	28.4	71.6	β
e	2.31	0.54	1.11	80.4(*h*)	9.1(*d*)		0.0	49.8	50.2	C-cap β
f	6.72	0.50	1.00	60.7(*k*)	36.3(*b*)		0.0	72.5	27.5	C-cap β
g	1.28	0.74	1.05	37.5(*h*)	28.0(*c*)	19.1(*o*)	6.9	83.8	9.3	mainly coil
h	2.35	0.62	1.04	62.4(*i*)	18.1(*j*)	10.2(*k*)	0.0	81.5	18.5	mainly coil
i	1.62	0.56	1.01	87.7(*a*)			0.0	94.5	5.5	mainly coil
j	0.96	1.03	1.01	17.0(*a*)	16.6(*b*)	16.1(*l*)	3.7	87.9	8.4	mainly coil
k	5.46	0.59	1.00	76.2(*l*)	13.6(*b*)		35.1	64.2	0.7	N-cap α
l	5.35	0.63	1.01	68.5(*m*)	9.2(*p*)	7.0(*c*)	44.4	54.9	0.7	N-cap α
m	30.04	0.43	6.74	33.8(*n*)	18.5(*p*)	9.7(*b*)	86.7	13.2	0.1	α
n	1.93	0.61	1.03	90.9(*o*)			68.4	31.3	0.3	C-cap α
o	2.60	0.60	1.02	74.7(*p*)	8.3(*m*)		43.1	56.8	0.1	C-cap α
p	3.41	0.46	1.00	58.1(*a*)	22.7(*c*)	11.1(*m*)	11.2	87.5	1.3	C-cap α to N-cap β

- PBm weist kleinen rmsd-wert auf (0.43 Å)
- überrepräsentiert: aliphatische AS
  unterrepräsentiert: α-Helixbrecher
- KLd zeigt breites Spektrum von spezifischen Stellen



Ergebnisse der Vorhersagestrategie am Beispiel des Proteins 2aak

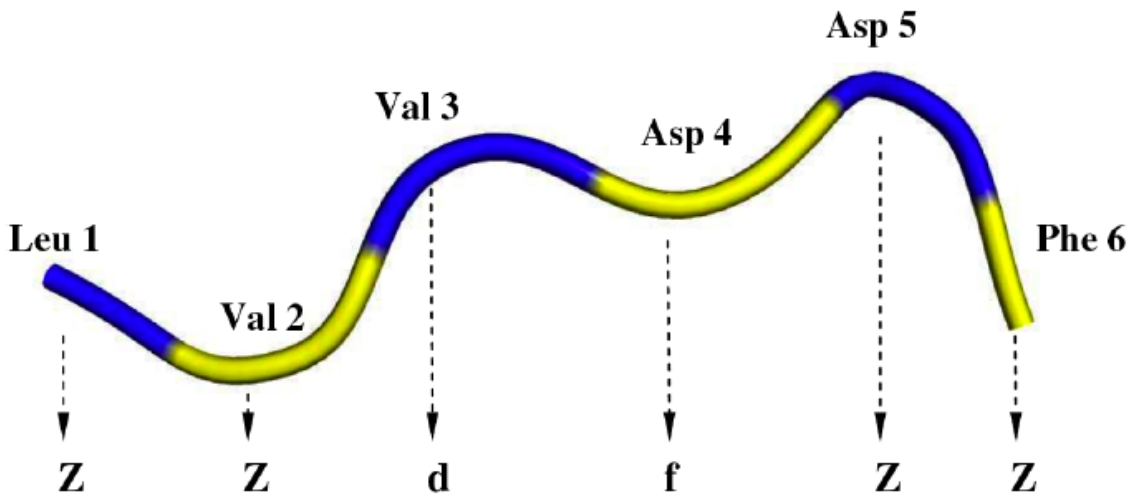
Sequence					Predicted PBs		
Left	c. window	Right	True PB	*Neq*	1st	2nd	3rd
DMSTP	**ARKLM**	RDFKR	*l*	2.74	*m*(11.21)	*l*(0.63)	*d*(0.36)
MSTPA	**RKLMR**	DFKRL	*m*	2.43	*m*(17.51)	*d*(0.74)	*f*(0.43)
STPAR	**KLMRD**	FKRLQ	*m*	2.06	*m*(33.68)	*f*(0.38)	*d*(0.18)
TPARK	**LMRDF**	KRLQQ	*m*	2.63	*m*(11.05)	*l*(0.51)	*k*(0.36)
PARKL	**MRDFK**	RLQQD	*m*	2.48	*m*(22.13)	*f*(1.25)	*b*(0.40)
ARKLM	**RDFKR**	LQQDP	*m*	3.78	*m*(7.77)	*k*(1.90)	*c*(0.54)
RKLMR	**DFKRL**	QQDPP	*m*	2.92	*m*(12.48)	*b*(0.94)	*c*(0.34)
KLMRD	**FKRLQ**	QDPPA	*m*	3.49	*m*(12.98)	*n*(2.60)	*p*(0.73)
LMRDF	**KRLQQ**	DPPAG	*m*	6.32	*m*(3.51)	*n*(0.55)	*d*(0.38)
MRDFK	**RLQQD**	PPAGI	*m*	8.61	*p*(2.02)	*b*(1.08)	*m*(1.02)
RDFKR	**LQQDP**	PAGIA	*m*	4.82	*p*(3.03)	*d*(1.12)	*c*(0.44)
DFKRL	**QQDPP**	AGIAG	*c*	2.55	*c*(4.43)	*d*(0.19)	*p*(0.12)
FKRLQ	**QDPPA**	GIAGA	*c*	3.10	*f*(13.43)	*c*(2.87)	*k*(0.23)
KRLQQ	**DPPAG**	IAGAG	*e*	5.45	*b*(7.14)	*e*(1.94)	*g*(1.68)
RLQQD	**PPAGI**	AGAGI	*h*	5.34	*b*(12.39)	*h*(6.72)	*l*(3.16)
LQQDP	**PAGIA**	GAGIS	*i*	4.97	*i*(11.29)	*p*(5.29)	*c*(1.40)
QQDPP	**AGIAG**	AGISG	*a*	4.75	*g*(15.58)	*a*(6.16)	*e*(3.24)
QDPPA	**GIAGA**	GISGA	*c*	6.75	*b*(7.15)	*h*(4.01)	*c*(2.32)

- Erstellen des Datensatzes:
  - Sequenzähnlichkeit $< 30\%$
  - Auflösung $< 2,5$ Å
  - Anzahl AS, die an Bindung beteiligt sind ≥ 3 AS

$\rightarrow$ Datensatz enthält 77 Bindungen in 70 Proteinen

- Erstellen des Datensatzes:
  - Sequenzähnlichkeit $< 30\%$
  - Auflösung $< 2,5$ Å
  - Anzahl AS, die an Bindung beteiligt sind ≥ 3 AS

  $\rightarrow$ Datensatz enthält 77 Bindungen in 70 Proteinen
- Umwandlung mittels PBE-Webinterface in Strukturalphabet



first-shell und second-shell Liganden

1^{st}-shell : Entfernung zwischen Metallion und Donor $\leq$ 2.5 Å

2^{nd}-shell : Entfernung $\leq$ 3.5 Å

first-shell und second-shell Liganden

1^{st}-shell : Entfernung zwischen Metallion und Donor $\leq$ 2.5 Å

2^{nd}-shell : Entfernung $\leq$ 3.5 Å

Strukturmotive

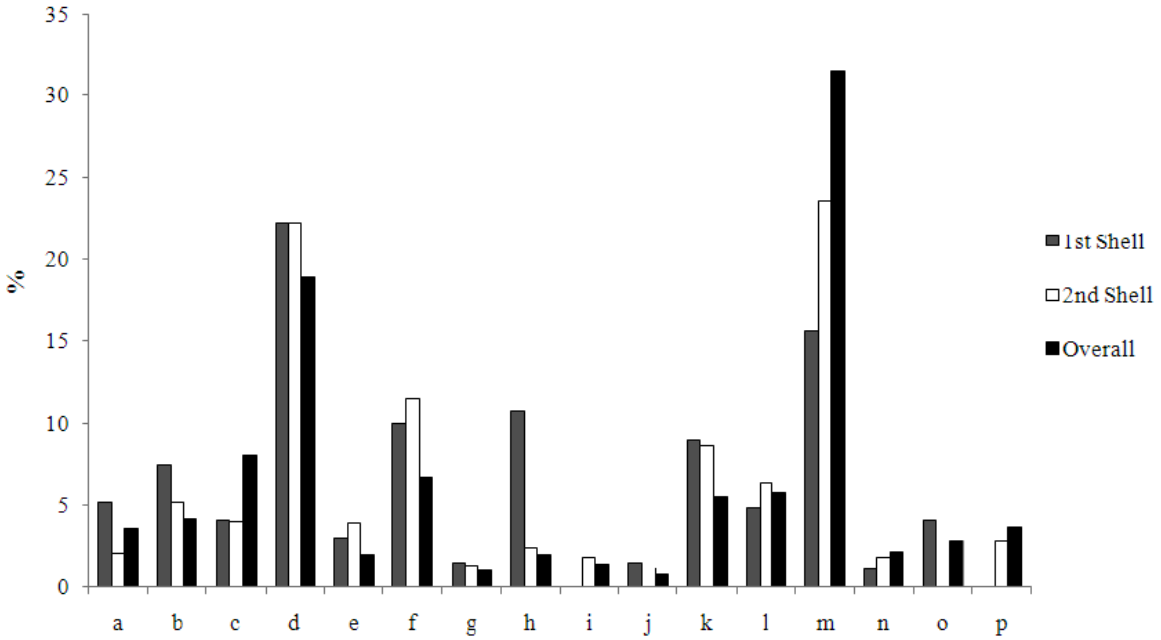
- $k \geq 3$; k beschreibt die Anzahl der Wiederholungen
- gleiche Strukturbuchstaben
- ähnlich große Zwischenräume

strukturelle Präferenzen der Mg^{2+}-Bindungstsellen

- Analyse der Struktur von 1^{st}- und 2^{nd} shell Bindungsstellen
- Vergleich der Frequenzen der beiden Schalen mit dem allgemeinen Auftreten der Struktur

strukturelle Präferenzen der Mg^{2+}-Bindungstsellen

- Analyse der Struktur von 1^{st}- und 2^{nd} shell Bindungsstellen
- Vergleich der Frequenzen der beiden Schalen mit dem allgemeinen Auftreten der Struktur



Mg^{2+}-Bindungsstellen präferieren Loops statt Helizes und Faltblätter

Suche nach strukturellen Motiven

1^{st}-shell : 4 Motive, die 21% aller Bindungsstellen repräsentieren

$\longrightarrow$ *e(24-47)h(24)k*

$\longrightarrow$ *f(1)h(109-349)b*

$\longrightarrow$ *f(2)h(126-158)m*

$\longrightarrow$ *k(26-29)h(1)a*

Suche nach strukturellen Motiven

1^{st}-shell : 4 Motive, die 21% aller Bindungsstellen repräsentieren

$\longrightarrow$ *e(24-47)h(24)k*

$\longrightarrow$ *f(1)h(109-349)b*

$\longrightarrow$ *f(2)h(126-158)m*

$\longrightarrow$ *k(26-29)h(1)a*

- gleiche Motive haben meist gleiche CATH-Nummern (hierarchische Klassifizierung von Proteindomänen)
- Möglichkeit für noch nicht klassifizierte Proteine?

Ergebnisse

Motif[a]	*PDB code*	*Mg^{2+}-Ligands*	*CATH number*[b]	*Functional Group*[c]	*EC code*[d]
e(24–47)h(24)k	1SJC	D^{189}, E^{214}, D^{239}	3.20.20.120	Lyase[e], Isomerase[f]	-
	1TKK	D^{191}, E^{219}, D^{244}	3.20.20.120	Isomerase[f]	-
	2AKZ	D^{244}, E^{292}, D^{317}	-	Lyase[e]	4.2.1.11
f(1)h(109–349)b	1O08	D^{1008}, D^{1010}, D^{1170}	3.40.50.1000	Isomerase[f]	5.4.2.6
	1U7P	D^{11}, D^{13}, D^{123}	NYC	Hydrolase[g]	-
	1WPG	D^{351}, T^{353}, D^{703}	3.40.50.1000	Hydrolase[g]	3.6.3.8
	2B82	D^{44}, D^{46}, D^{167}	3.40.50.1000	Hydrolase[g]	3.1.3.2
	2C4N	D^{9}, D^{11}, D^{201}	NYC	Hydrolase[g]	-
f(2)h(126–158)m	1KA1	D^{142}, D^{145}, D^{294}	3.30.540.10	Hydrolase[g]	3.1.3.7
	1NUY	D^{1118}, D^{1121}, E^{1280}	3.30.540.10+ 3.40.190.80	Hydrolase[g]	3.1.3.11
	2BJI	E^{1090}, D^{1093}, D^{1220}	3.30.540.10+ 3.40.190.80	Hydrolase[g]	3.1.3.25
k(26–29)h(1)a	1ITZ	D^{168}, N^{198}, I^{200}	3.40.50.970	Transferase[h]	2.2.1.1
	1POX	D^{447}, N^{474}, Q^{476}	3.40.50.970+ 3.40.50.1220	Oxidoreductase[i]	1.2.3.3
	1UMD	D^{175}, N^{204}, Y^{206}	3.40.50.970	Oxidoreductase[i]	1.2.4.4
	1ZPD	D^{440}, N^{467}, G^{469}	3.40.50.970	Lyase[e]	4.1.1.1
	2C3M	D^{963}, T^{991}, V^{993}	3.40.50.970	Oxidoreductase[i]	1.2.7.1

Wie spezifisch sind die Resultate?

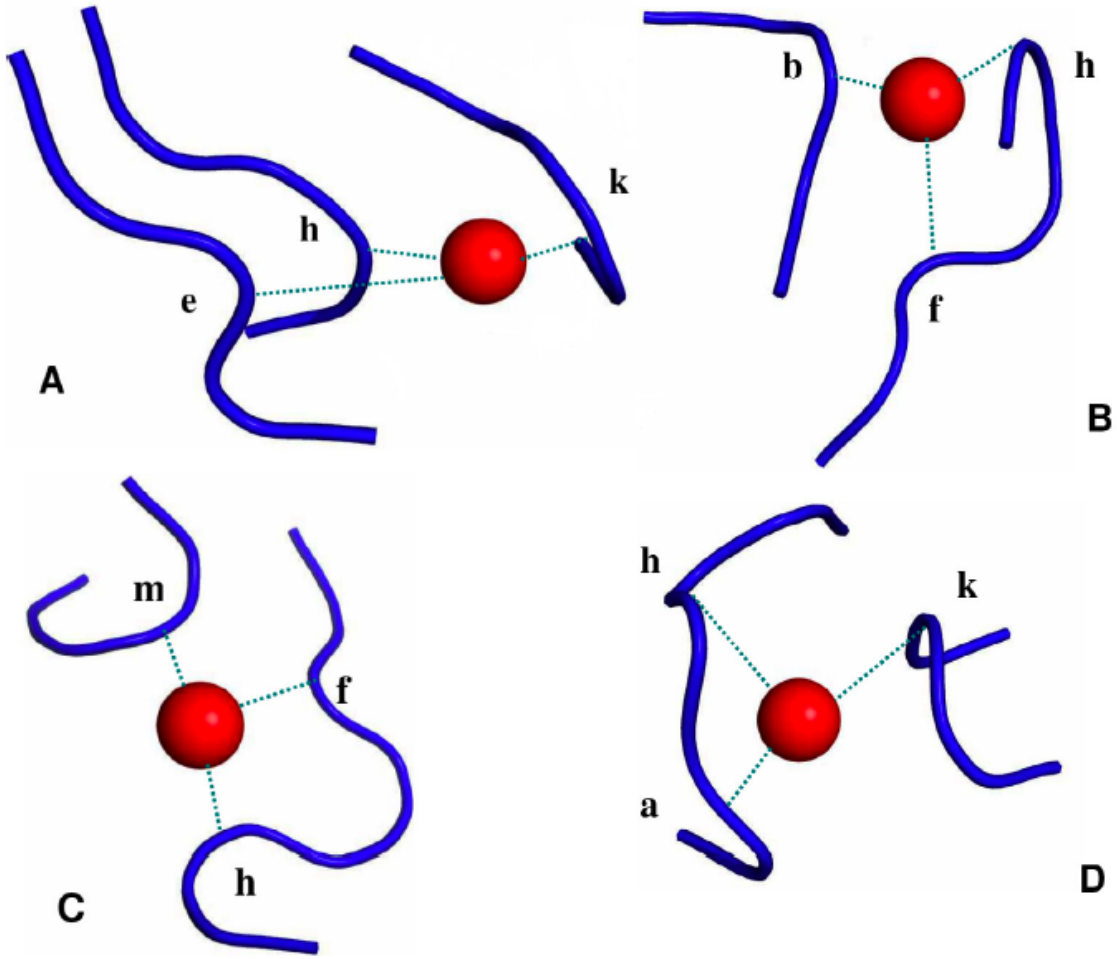
- Abgleich gegen Nichtmetallproteine aus de Brevern's DB (Auswahl der Proteine nach vorab definierten Kriterien)

  → matches für *f(1)h(109-349)b* und *k(26-29)h(1)a*
  → *e(24-47)h(24)k* & *f(2)h(126-158)m* sind metallspezifisch

Wie spezifisch sind die Resultate?

- Abgleich gegen Nichtmetallproteine aus de Brevern's DB (Auswahl der Proteine nach vorab definierten Kriterien)
  - → matches für *f(1)h(109-349)b* und *k(26-29)h(1)a*
  - → *e(24-47)h(24)k* & *f(2)h(126-158)m* sind metallspezifisch
- Abgleich gegen calciumbindende Proteine (Anwendung des gleichen Verfahrens wie für die Mg^{2+}-Bindungsstellen)
  - → *f(1)h(109-349)b* und *k(26-29)h(1)a* wurden in 1 bzw. 2 Proteinen gefunden
  - → *e(24-47)h(24)k* & *f(2)h(126-158)m* sind metall- und magnesiumspezifisch

4 extrahierte Bindungsmotive :



Literatur

- Minko Dudev, Carmay Lim *Discovering structurel motifs using a structural alphabet: Application to magnesium-binding sites*, Bioinformatics 2007 Mar 28; 8(106)
- de Brevern AG, Etchebest C, Hazout S *Bayesian Probabilistic Approach for Predicting Backbone Sturctures in Terms of Protein Blocks* PROTEINS: Structures, Function and Genetics 2000