Context-specific Independence Mixture Modelling for Protein Families

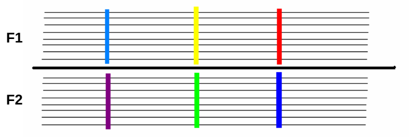
Benjamin Georgi
Jörg Schultz
Alexander Schliep

Introduction

- Protein families fall into sub families of similar but different function
- Specific function of sub families is often determined by a small number of residues
  → functional residues
- Example: Malate / Lactate dehydrogenase – single residue determines specificity

Functional Positions

- Multiple sequence alignment (MSA) :



- Two sub families, three functional positions
- Strong signal of subgroup specific conservation

Introduction

- **Problem:** Clustering of protein families and simultaneous prediction of functional residues
- Prior approaches:
  - Mostly supervised, requiring additional prior knowledge
  - Mostly based on phylogenetic trees
- Our approach: First unsupervised method that does not require a tree

→ Context-specific independence (CSI) mixture models

Mixture Models

- Given random variable $X = (X_1, ..., X_p)$
- X represents row in a MSA
- K component mixture density:

$$P(x\,|\,\Theta) = \sum_{i=1}^{K} w_i f_i(x, \theta_i)$$

s.t. $\sum_{i=1}^{K} w_i = 1,\ w_i \geq 0 \qquad \Theta = (\{\mathrm{w}_\mathrm{i}\}_{\mathrm{i=1..K}}, \{\theta_\mathrm{i}\}_{\mathrm{i=1..K}})$

Mixture Models

Example: MSA lenght 4 $X = (X_1, X_2, X_3, X_4)$
4 component mixture

$$\sum_{i=1}^{4} w_i f_i(X, \theta_i) =$$

$$\begin{aligned}
& w_1\, f_{11}(X_1, \theta_{11})\, f_{12}(X_2, \theta_{12})\, f_{13}(X_3, \theta_{13})\, f_{14}(X_4, \theta_{14}) + \\
& w_2\, f_{21}(X_1, \theta_{21})\, f_{22}(X_2, \theta_{22})\, f_{23}(X_3, \theta_{23})\, f_{24}(X_4, \theta_{24}) + \\
& w_3\, f_{31}(X_1, \theta_{31})\, f_{32}(X_2, \theta_{32})\, f_{33}(X_3, \theta_{33})\, f_{34}(X_4, \theta_{34}) + \\
& w_4\, f_{41}(X_1, \theta_{41})\, f_{42}(X_2, \theta_{42})\, f_{43}(X_3, \theta_{43})\, f_{44}(X_4, \theta_{44})
\end{aligned}$$

Mixture Models

Model parameterization

$$\begin{array}{cc|c|c|c|c|}
 & & X_1 & X_2 & X_3 & X_4 \\
\hline
C_1 & w_1 & \theta_{11} & \theta_{12} & \theta_{13} & \theta_{14} \\
\hline
C_2 & w_2 & \theta_{21} & \theta_{22} & \theta_{23} & \theta_{24} \\
\hline
C_3 & w_3 & \theta_{31} & \theta_{32} & \theta_{33} & \theta_{34} \\
\hline
C_4 & w_4 & \theta_{41} & \theta_{42} & \theta_{43} & \theta_{44} \\
\hline
\end{array}$$

Mixture Models

Model structure matrix:

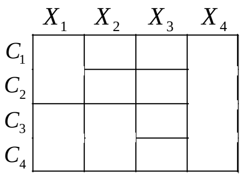
one atomar distribution per feature and component

	X_1	X_2	X_3	X_4
C_1				
C_2				
C_3				
C_4				

CSI Mixture Models

Model structure matrix:

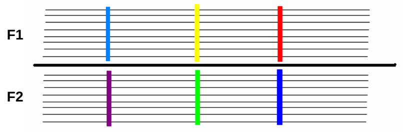
variable number of parameters for each feature



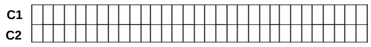
⟶ use model structure to predict functional positions

CSI for Protein Families

- MSA

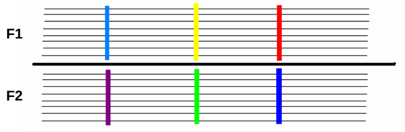


- Conventional mixture

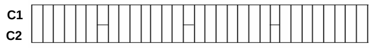


CSI for Protein Families

- MSA

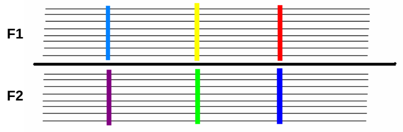


- CSI Structure matrix (idealized)

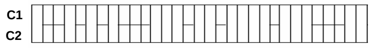


CSI for Protein Families

- MSA



- CSI Structure matrix (more realistic)



Prediction of Functional Residues

- Rank features by importance for characterization of a cluster
- Ranking of feature *j* for component *i* by:

$$KL_{sym}(\theta_{ij},\theta_0) = \frac{KL(\theta_{ij},\theta_0) + KL(\theta_0,\theta_{ij})}{2}$$

- Take highest ranking features as putative functional residues

CSI Structure Learning

- Question: How to learn a CSI structure from data?
- Bayesian approach: Score models by posterior distribution

$$P(M \mid D) \propto P(D \mid M) P(M)$$

CSI Structure Learning

- Model posterior

$$\boxed{P(M \mid D)} = \boxed{\prod_{i=1}^{N} P(x_i \mid \Theta)}\ \boxed{P(\Theta)}\ \boxed{P(M)}$$

Model posterior

Likelihood

Parameter prior

Structure prior

CSI Structure Learning

- Model posterior

$$\boxed{P(M \mid D)} = \boxed{\prod_{i=1}^{N} P(x_i \mid \Theta)}\ \boxed{P(\Theta)}\ \boxed{P(M)}$$

Model posterior

Likelihood

Parameter prior

Structure prior

Criterion used to select structure and parameter estimates

CSI Structure Learning

- Model posterior

$$\boxed{P(M \mid D)} = \boxed{\prod_{i=1}^{N} P(x_i \mid \Theta)}\ \boxed{P(\Theta)}\ \boxed{P(M)}$$

Model posterior

Likelihood

Parameter prior

Structure prior

Probability of the data under the mixture model

CSI Structure Learning

- Model posterior

$$\boxed{P(M \mid D)} = \boxed{\prod_{i=1}^{N} P(x_i \mid \Theta)}\ \boxed{P(\Theta)}\ \boxed{P(M)}$$

Model posterior

Likelihood

Parameter prior

Structure prior

Typically uninformative prior, acts as pseudo counts, conjugate Dirchlet distribution

CSI Structure Learning

- Model posterior

$$\boxed{P(M \mid D)} = \boxed{\prod_{i=1}^{N} P(x_i \mid \Theta)}\ \boxed{P(\Theta)}\ \boxed{P(M)}$$

Model posterior

Likelihood

Parameter prior

Structure prior

Introduces preference for simpler model, acts as a regularizer, simple factored form

Learning Algorithm

- Structural EM framework (Friedman 1998)
- Efficiently score candidate structures based on the expected sufficent statistics
- Perform greedy search over structure space to arrive at final structure

Mixtures for Proteins

- Using sequence to infer structural properties of the proteins
- Different amino acid substitutions have different structural impacts
- Need notion of amino acid similarity in the model to guide the structure learning
- Remark: related to substitution matrices in phylogeny

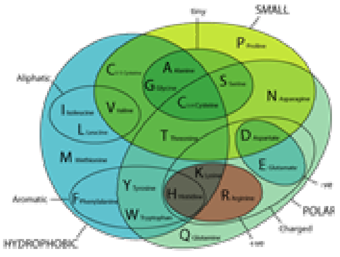
Conceptional problem

- Example:

```
      L E
      L E
C1    L E
      L E
      -----
      N D
C2    N D
      N D
      N D
```

- Aspartate (D) and Glutamate (E) much more similar than Leucine (L) and Asparagine (N)
- E/D column might best be modeled with a single distribution in the structure

Amino acid properties

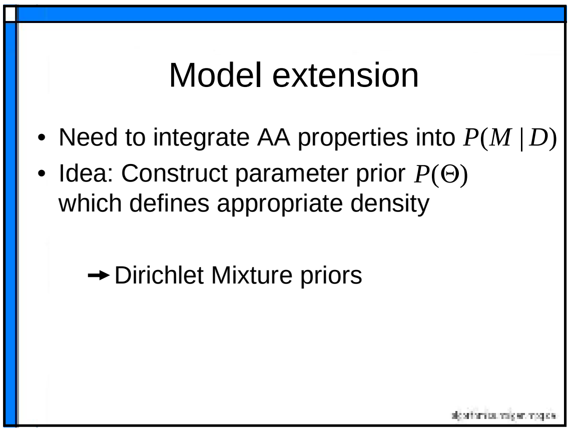


Livingstone and Barton (1993)

Amino acid properties

- **AA Property table:**

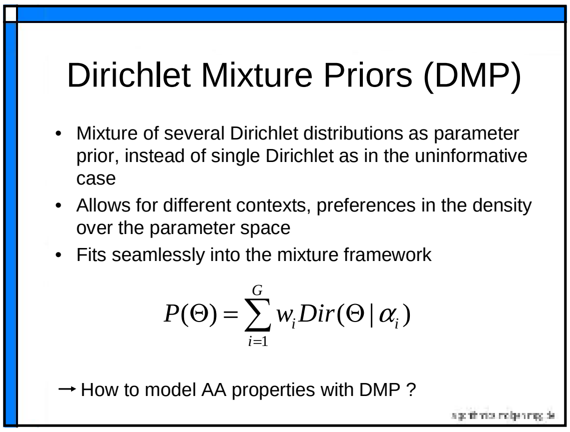
I	L	V	C	A	G	M	F	Y	W	H	K	R	E	Q	D	N	S	T	P	
X	X	X	X	X	X	X	X	X	X	X	·	·	·	·	·	·	·	X	·	Hydrophobic
·	·	·	·	·	·	·	·	X	X	X	X	X	X	X	X	X	X	·	X	Polar
·	·	X	X	X	X	·	·	·	·	·	·	·	·	·	X	X	X	X	X	Small
·	·	·	·	X	X	·	·	·	·	·	·	·	·	·	·	·	X	·	·	Tiny
X	X	X	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	Aliphatic
·	·	·	·	·	·	·	X	X	X	X	·	·	·	·	·	·	·	·	·	Aromatic
·	·	·	·	·	·	·	·	·	·	X	X	X	·	·	·	·	·	·	·	Positive
·	·	·	·	·	·	·	·	·	·	·	·	·	X	·	X	·	·	·	·	Negative
·	·	·	·	·	·	·	·	·	·	X	X	X	X	·	X	·	·	·	·	Charged



Model extension

- Need to integrate AA properties into $P(M \mid D)$
- Idea: Construct parameter prior $P(\Theta)$ which defines appropriate density

→ Dirichlet Mixture priors

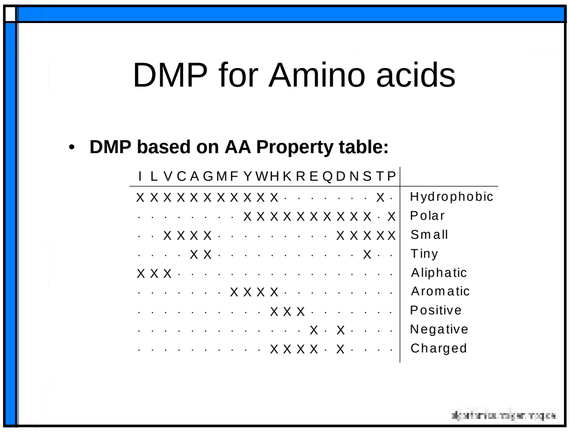


Dirichlet Mixture Priors (DMP)

- Mixture of several Dirichlet distributions as parameter prior, instead of single Dirichlet as in the uninformative case
- Allows for different contexts, preferences in the density over the parameter space
- Fits seamlessly into the mixture framework

$$P(\Theta) = \sum_{i=1}^{G} w_i Dir(\Theta \mid \alpha_i)$$

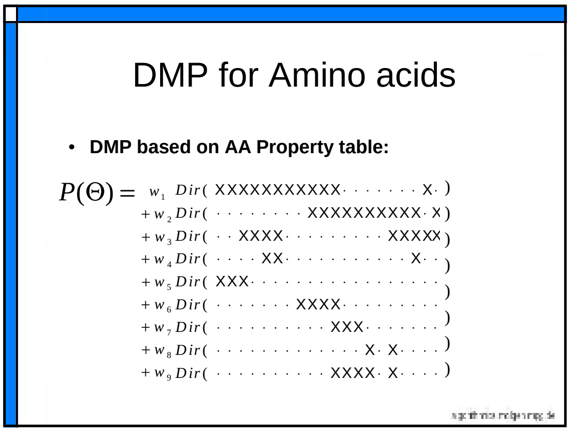
→ How to model AA properties with DMP ?



DMP for Amino acids

- **DMP based on AA Property table:**

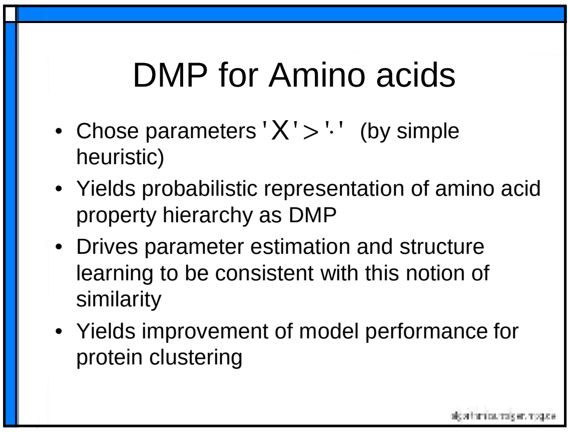
I	L	V	C	A	G	M	F	Y	W	H	K	R	E	Q	D	N	S	T	P	
X	X	X	X	X	X	X	X	X	X	X	·	·	·	·	·	·	·	X	·	Hydrophobic
·	·	·	·	·	·	·	·	X	X	X	X	X	X	X	X	X	X	·	X	Polar
·	·	X	X	X	X	·	·	·	·	·	·	·	·	·	X	X	X	X	X	Small
·	·	·	·	X	X	·	·	·	·	·	·	·	·	·	·	·	X	·	·	Tiny
X	X	X	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	Aliphatic
·	·	·	·	·	·	·	X	X	X	X	·	·	·	·	·	·	·	·	·	Aromatic
·	·	·	·	·	·	·	·	·	·	X	X	X	·	·	·	·	·	·	·	Positive
·	·	·	·	·	·	·	·	·	·	·	·	·	X	·	X	·	·	·	·	Negative
·	·	·	·	·	·	·	·	·	·	X	X	X	X	·	X	·	·	·	·	Charged



DMP for Amino acids

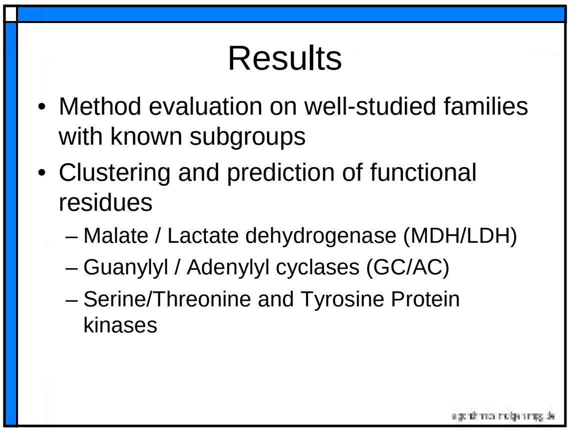
- **DMP based on AA Property table:**

$$\begin{aligned} P(\Theta) = {} & w_1\, Dir(\mathrm{XXXXXXXXXXX}\cdot\cdot\cdot\cdot\cdot\cdot\cdot\mathrm{X}\cdot) \\ & + w_2\, Dir(\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\mathrm{XXXXXXXXXX}\cdot\mathrm{X}) \\ & + w_3\, Dir(\cdot\cdot\mathrm{XXXX}\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\mathrm{XXXXX}) \\ & + w_4\, Dir(\cdot\cdot\cdot\cdot\mathrm{XX}\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\mathrm{X}\cdot\cdot) \\ & + w_5\, Dir(\mathrm{XXX}\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot) \\ & + w_6\, Dir(\cdot\cdot\cdot\cdot\cdot\cdot\cdot\mathrm{XXXX}\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot) \\ & + w_7\, Dir(\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\mathrm{XXX}\cdot\cdot\cdot\cdot\cdot\cdot\cdot) \\ & + w_8\, Dir(\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\mathrm{X}\cdot\mathrm{X}\cdot\cdot\cdot\cdot) \\ & + w_9\, Dir(\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\mathrm{XXXX}\cdot\mathrm{X}\cdot\cdot\cdot\cdot) \end{aligned}$$



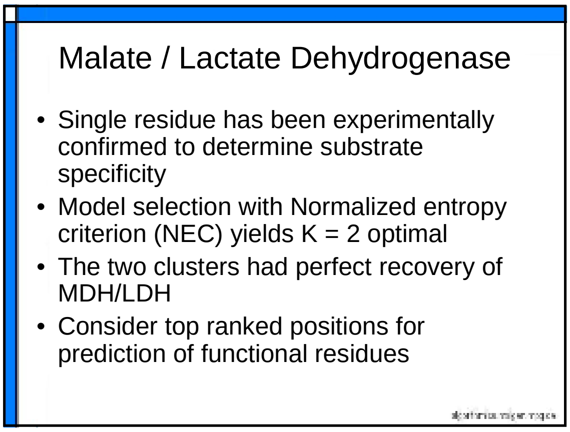
DMP for Amino acids

- Chose parameters 'X' > '·' (by simple heuristic)
- Yields probabilistic representation of amino acid property hierarchy as DMP
- Drives parameter estimation and structure learning to be consistent with this notion of similarity
- Yields improvement of model performance for protein clustering



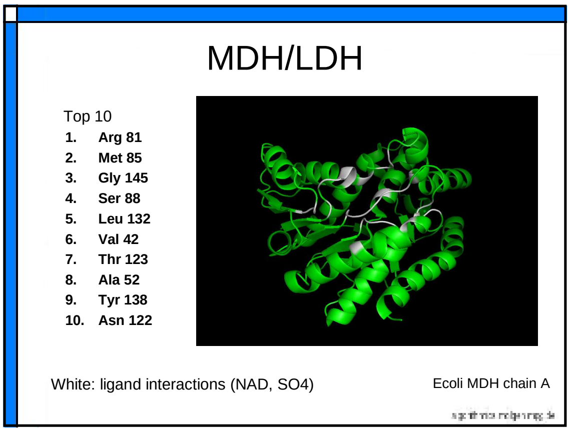
Results

- Method evaluation on well-studied families with known subgroups
- Clustering and prediction of functional residues
  - Malate / Lactate dehydrogenase (MDH/LDH)
  - Guanylyl / Adenylyl cyclases (GC/AC)
  - Serine/Threonine and Tyrosine Protein kinases



Malate / Lactate Dehydrogenase

- Single residue has been experimentally confirmed to determine substrate specificity
- Model selection with Normalized entropy criterion (NEC) yields K = 2 optimal
- The two clusters had perfect recovery of MDH/LDH
- Consider top ranked positions for prediction of functional residues



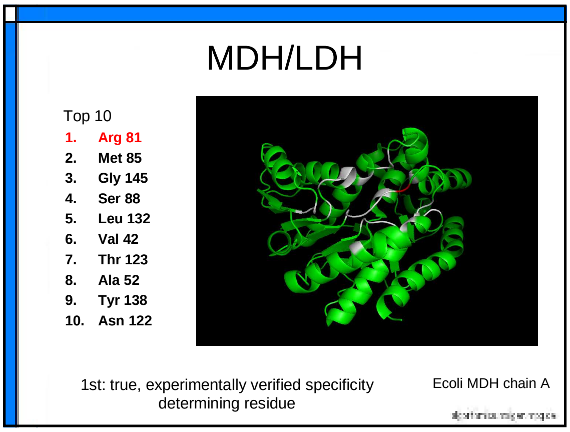
MDH/LDH

Top 10

1. **Arg 81**
2. **Met 85**
3. **Gly 145**
4. **Ser 88**
5. **Leu 132**
6. **Val 42**
7. **Thr 123**
8. **Ala 52**
9. **Tyr 138**
10. **Asn 122**

White: ligand interactions (NAD, SO4)

Ecoli MDH chain A



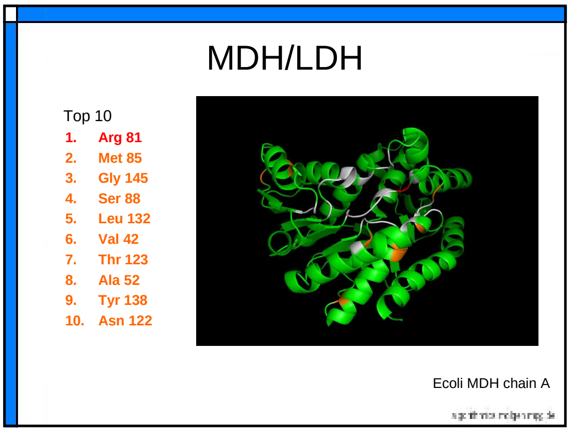
MDH/LDH

Top 10

1. **Arg 81**
2. **Met 85**
3. **Gly 145**
4. **Ser 88**
5. **Leu 132**
6. **Val 42**
7. **Thr 123**
8. **Ala 52**
9. **Tyr 138**
10. **Asn 122**

1st: true, experimentally verified specificity determining residue

Ecoli MDH chain A

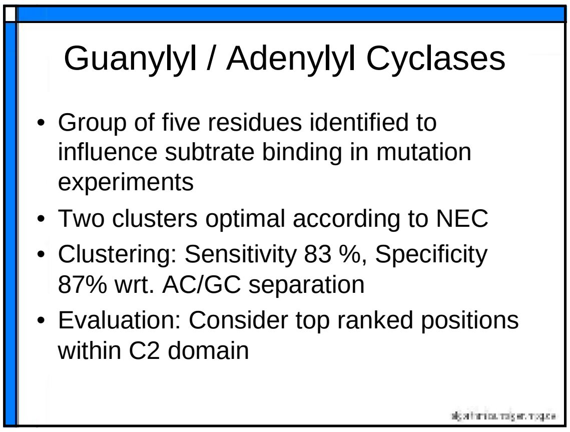


MDH/LDH

Top 10

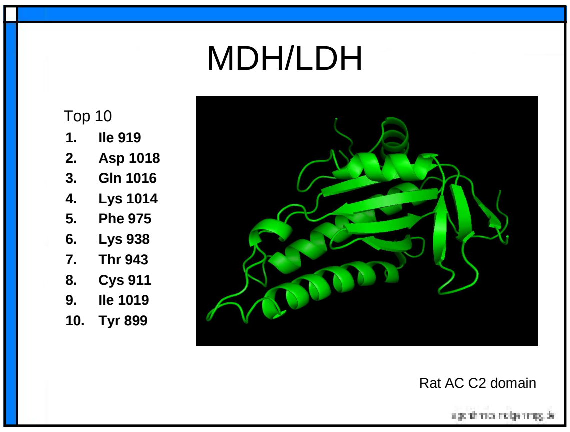
1. **Arg 81**
2. **Met 85**
3. **Gly 145**
4. **Ser 88**
5. **Leu 132**
6. **Val 42**
7. **Thr 123**
8. **Ala 52**
9. **Tyr 138**
10. **Asn 122**

Ecoli MDH chain A



Guanylyl / Adenylyl Cyclases

- Group of five residues identified to influence subtrate binding in mutation experiments
- Two clusters optimal according to NEC
- Clustering: Sensitivity 83 %, Specificity 87% wrt. AC/GC separation
- Evaluation: Consider top ranked positions within C2 domain

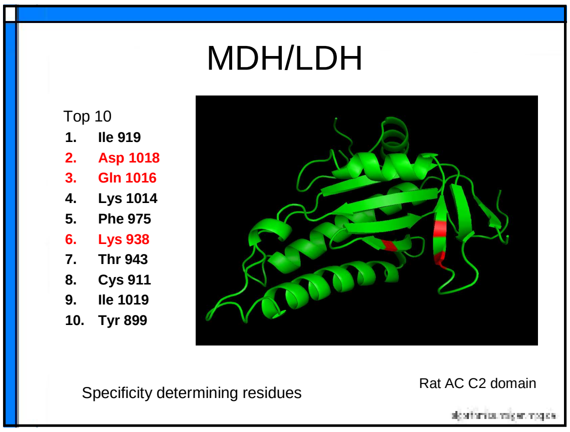


MDH/LDH

Top 10

1. **Ile 919**
2. **Asp 1018**
3. **Gln 1016**
4. **Lys 1014**
5. **Phe 975**
6. **Lys 938**
7. **Thr 943**
8. **Cys 911**
9. **Ile 1019**
10. **Tyr 899**

Rat AC C2 domain



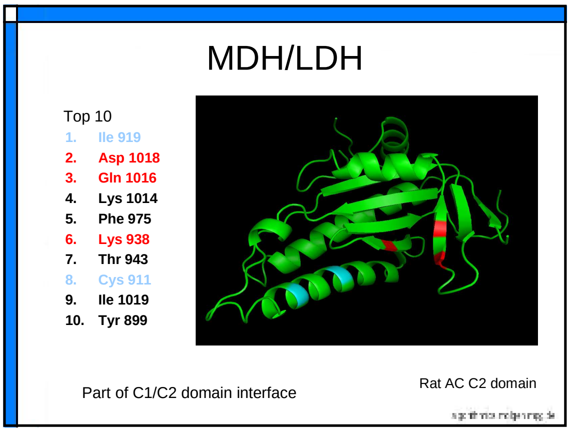
MDH/LDH

Top 10

1. Ile 919
2. Asp 1018
3. Gln 1016
4. Lys 1014
5. Phe 975
6. Lys 938
7. Thr 943
8. Cys 911
9. Ile 1019
10. Tyr 899

Specificity determining residues

Rat AC C2 domain



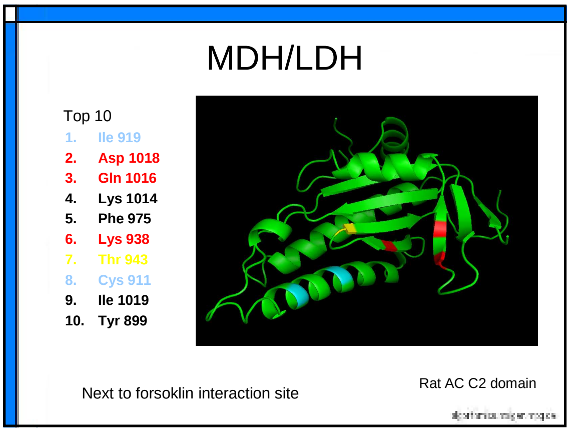
MDH/LDH

Top 10

1. Ile 919
2. Asp 1018
3. Gln 1016
4. Lys 1014
5. Phe 975
6. Lys 938
7. Thr 943
8. Cys 911
9. Ile 1019
10. Tyr 899

Part of C1/C2 domain interface

Rat AC C2 domain



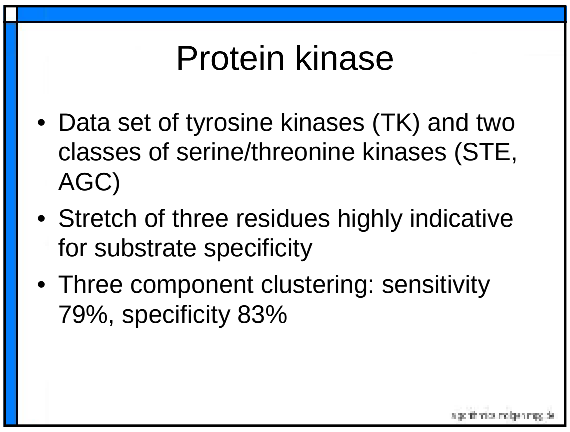
MDH/LDH

Top 10

1. Ile 919
2. Asp 1018
3. Gln 1016
4. Lys 1014
5. Phe 975
6. Lys 938
7. Thr 943
8. Cys 911
9. Ile 1019
10. Tyr 899

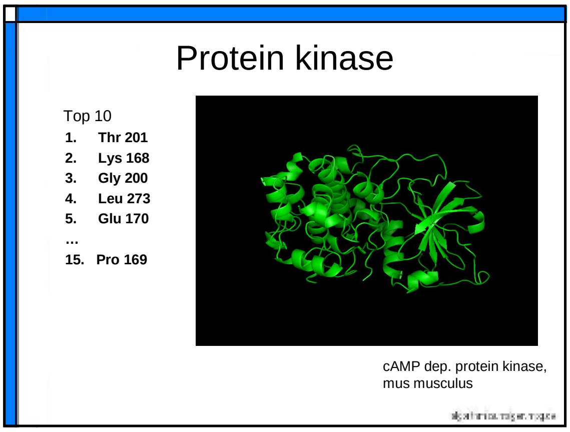
Next to forsoklin interaction site

Rat AC C2 domain



Protein kinase

- Data set of tyrosine kinases (TK) and two classes of serine/threonine kinases (STE, AGC)
- Stretch of three residues highly indicative for substrate specificity
- Three component clustering: sensitivity 79%, specificity 83%



Protein kinase

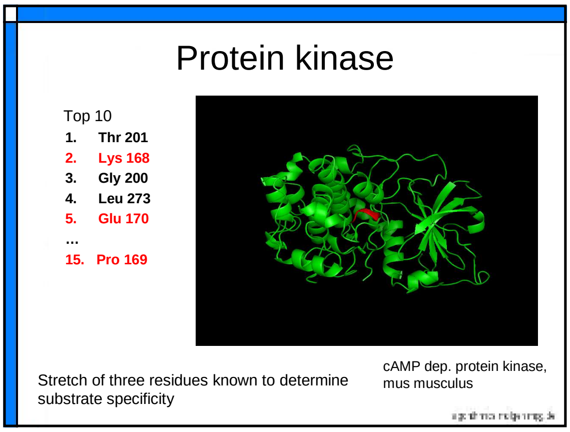
Top 10

1. Thr 201
2. Lys 168
3. Gly 200
4. Leu 273
5. Glu 170

…

15. Pro 169

cAMP dep. protein kinase, mus musculus



Protein kinase

Top 10

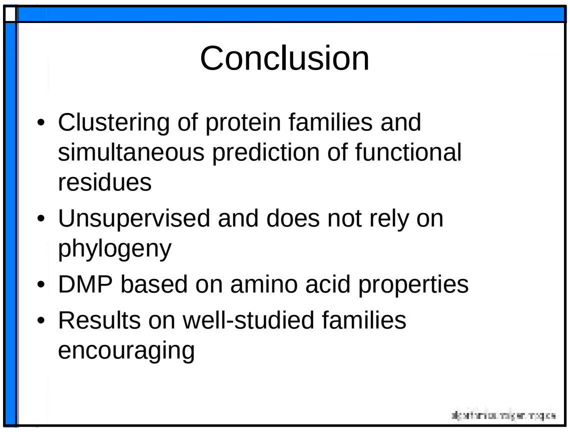
1. Thr 201
2. Lys 168
3. Gly 200
4. Leu 273
5. Glu 170

…

15. Pro 169

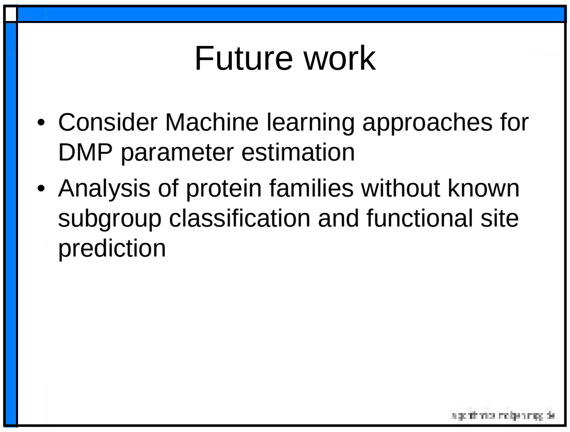
Stretch of three residues known to determine substrate specificity

cAMP dep. protein kinase, mus musculus



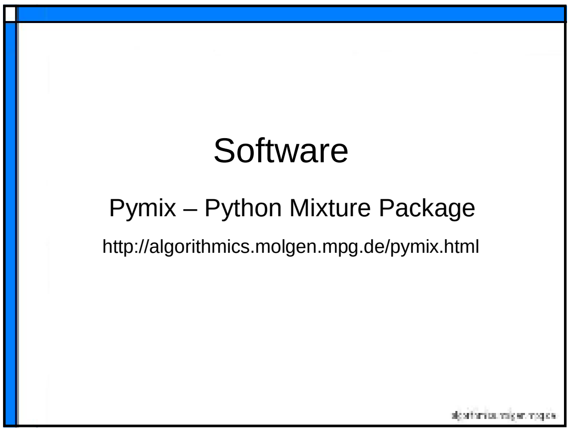
Conclusion

- Clustering of protein families and simultaneous prediction of functional residues
- Unsupervised and does not rely on phylogeny
- DMP based on amino acid properties
- Results on well-studied families encouraging



Future work

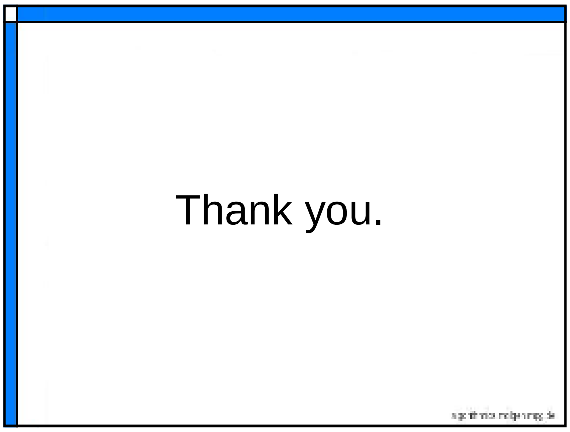
- Consider Machine learning approaches for DMP parameter estimation
- Analysis of protein families without known subgroup classification and functional site prediction



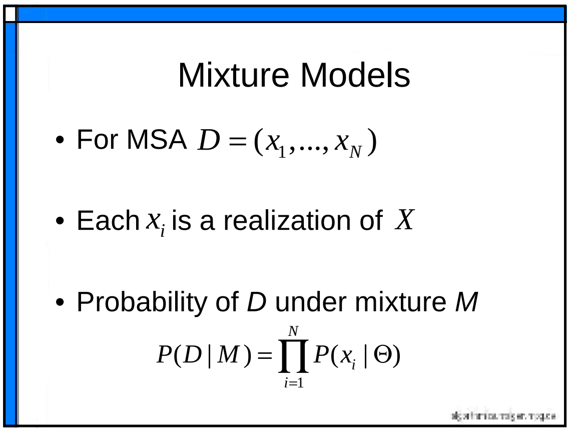
Software

Pymix – Python Mixture Package

http://algorithmics.molgen.mpg.de/pymix.html



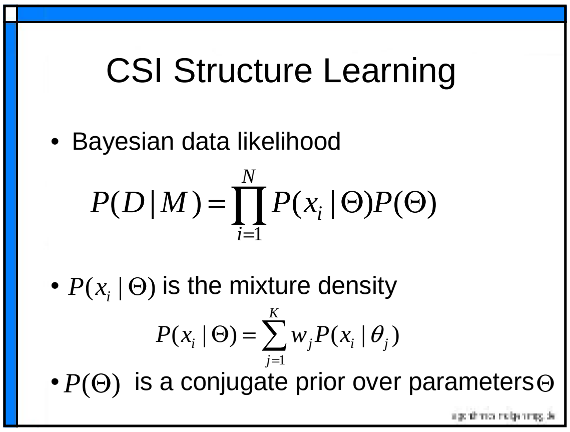
Thank you.



Mixture Models

- For MSA $D = (x_1, \ldots, x_N)$
- Each x_i is a realization of X
- Probability of *D* under mixture *M*

$$P(D|M) = \prod_{i=1}^{N} P(x_i | \Theta)$$



CSI Structure Learning

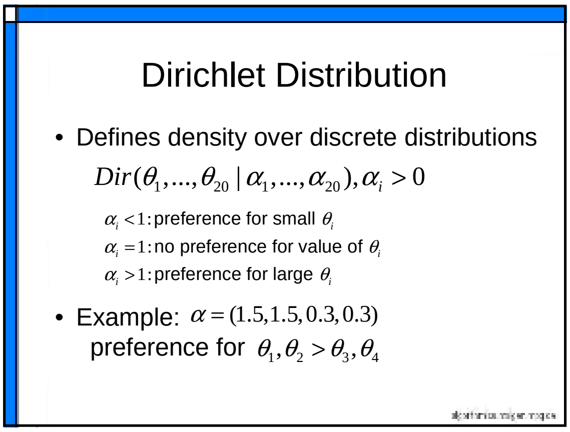
- Bayesian data likelihood

$$P(D|M) = \prod_{i=1}^{N} P(x_i | \Theta) P(\Theta)$$

- $P(x_i | \Theta)$ is the mixture density

$$P(x_i | \Theta) = \sum_{j=1}^{K} w_j P(x_i | \theta_j)$$

- $P(\Theta)$ is a conjugate prior over parameters Θ



Dirichlet Distribution

- Defines density over discrete distributions

  $Dir(\theta_1,\ldots,\theta_{20} \mid \alpha_1,\ldots,\alpha_{20}), \alpha_i > 0$

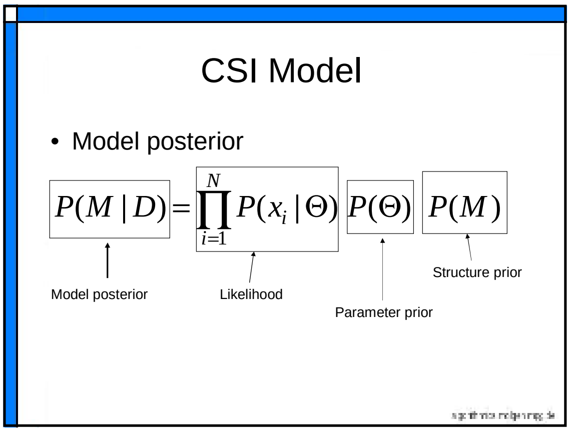
  $\alpha_i < 1$: preference for small θ_i

  $\alpha_i = 1$: no preference for value of θ_i

  $\alpha_i > 1$: preference for large θ_i

- Example: $\alpha = (1.5, 1.5, 0.3, 0.3)$

  preference for $\theta_1, \theta_2 > \theta_3, \theta_4$



CSI Model

- Model posterior

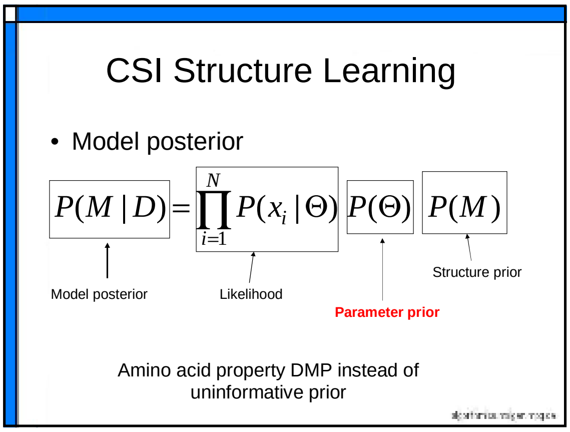
$$\boxed{P(M \mid D)} = \boxed{\prod_{i=1}^{N} P(x_i \mid \Theta)}\ \boxed{P(\Theta)}\ \boxed{P(M)}$$

Model posterior

Likelihood

Parameter prior

Structure prior



CSI Structure Learning

- Model posterior

$$\boxed{P(M \mid D)} = \boxed{\prod_{i=1}^{N} P(x_i \mid \Theta)}\ \boxed{P(\Theta)}\ \boxed{P(M)}$$

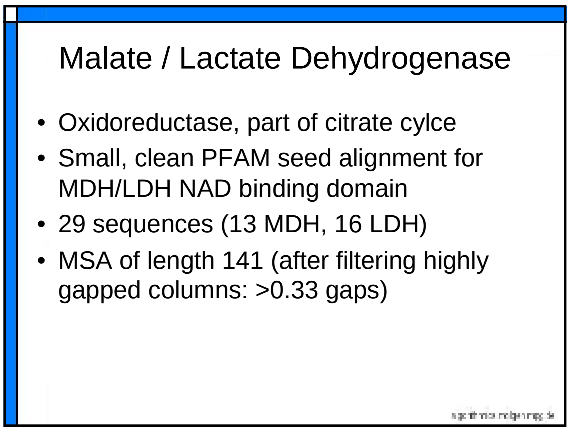
Model posterior

Likelihood

Parameter prior

Structure prior

Amino acid property DMP instead of
uninformative prior



Malate / Lactate Dehydrogenase

- Oxidoreductase, part of citrate cylce
- Small, clean PFAM seed alignment for MDH/LDH NAD binding domain
- 29 sequences (13 MDH, 16 LDH)
- MSA of length 141 (after filtering highly gapped columns: >0.33 gaps)