Context-Specific Independence Mixture Modelling for Protein Families

Benjamin Georgi [1], Jörg Schultz [2], and Alexander Schliep [1]

[1] Max Planck Institute for Molecular Genetics, Dept. of Computational Molecular Biology, Ihnestrasse 73, 14195 Berlin, Germany
[2] Universität Würzburg, Dept. of Bioinformatics, 97074 Wuerzburg, Germany

Abstract. Protein families can be divided into subgroups with functional differences. The analysis of these subgroups and the determination of which residues convey substrate specificity is a central question in the study of these families. We present a clustering procedure using the *context-specific independence* mixture framework using a Dirichlet mixture prior for simultaneous inference of subgroups and prediction of specificity determining residues based on multiple sequence alignments of protein families. Application of the method on several well studied families revealed a good clustering performance and ample biological support for the predicted positions. The software we developed to carry out this analysis *PyMix - the Python mixture package* is available from *http://www.algorithmics.molgen.mpg.de/pymix.html*.

1 Introduction

Proteins within the same family commonly fall into sub categories which differ by functional specificity. The categorization and analysis of these subgroups is one of the central challenges in the study of these families. In particular it is of interest which residues determine functional specificity of a subgroup. These functional residues are characterized by a strong signal of subgroup specific conservation.

A number of studies have focused on the question how to detect residues which determine functional specificity based on prior knowledge of subtype membership. A review of these methods can be found in [14]. Among the approaches taken were relative entropy based scores [12], classification based on similarity to a data base of functional residue templates [4], contrasting position specific conservation in orthologues and paralogues to predict functional residues [21]. In [26] the authors use known reference protein 3D structures to find conserved discriminatory surface residues. One major limitation of these *supervised* approaches is the requirement of biological expert annotation of the number of subtypes and subtype assignments for each sequence. Which then limits usefulness of these methods to cases where prior biological knowledge is abundant. In the absence of such knowledge the inference of the subgroups becomes one central aspect of the prediction of functional residues. In many cases the subgroup structure of a given family is a direct consequence of evolutionary divergence of homologue sequences. As such it is not surprising that methods based on

J.N. Kok et al. (Eds.): PKDD 2007, LNAI 4702, pp. 79–90, 2007.

the phylogenetic tree of a family have been extensively and successfully used to study protein family subgroups [15,16,22,25]. However, the performance of these methods does degrade in cases where the evolutionary divergence between subgroups is large. Moreover phylogeny does not account for situations where functional relatedness of proteins arose from a process of convergent evolution. As such there is a need for additional methods for detection and analysis of the subgroups inherent in a set of related sequences. Here, we present the first unsupervised approach to simultaneously cluster related sequences and predict functional residues which does not rely on a phylogenetic tree. Prior work either relies on inference of phylogenetic trees or is unsupervised.

The clustering procedure employs the Bayesian *context-specific independence* mixture framework [9]. CSI mixtures have for instance been used for modeling of transcription factor binding sites [9], clustering of gene expression data [1] or the analysis of complex genetic diseases [10]. The central idea of the *context-specific independence* model is to adapt the number of model parameters to a level which is appropriate for a given data set. This notion of automatic adaption of a probabilistic model to the data has received considerable attention in the context of Bayesian networks [3,5,7].

One of the challenges of clustering protein families into subgroups based on the sequence is that the discriminating features one attempts to learn are a property of the structure rather than the sequence. As an example, consider three subgroups with perfect conservation of amino acids Leucine, Isoleucine and Tryptophan respectively at one position. A naive application of a clustering would consider said position to be highly discriminative for all three groups. Of course, this would be misleading due to the great similarity in chemical properties between Leucine and Isoleucine which makes them, to some extent, synonymous as far as structure is concerned. To adapt the CSI mixture model for this situation we apply a parameter prior in form of a mixture of Dirichlet distributions. These Dirichlet mixture priors have been successfully used to improve generalization properties of parameter estimates for probabilistic models for small sample sizes [23]. In the CSI framework a suitably chosen prior additionally acts to guide the structure learning towards distributions indicative of structural differences between the subgroups.

2 Methods

2.1 CSI Mixture Models

In this section we briefly introduce notation for conventional mixture models and our extension in the *context-specific independence* framework. For a more in depth coverage the reader is referred to [20] and [9] respectively. Let $X_1, ..., X_p$ be discrete random variables over the 20 amino acids and a gap symbol representing a multiple sequence alignment (MSA) with p positions (see Fig. 1a for an example). Given a data set D of N realizations, $D = x_1, ..., x_N$ with $x_i = (x_{i1}, ..., x_{ip})$ a conventional mixture density is given by

$$P(x_i) = \sum_{k=1}^{K} \pi_k \; f_k(x_i|\theta_k), \tag{1}$$

where the π_k are the mixture coefficients, $\sum_{k=1}^{K} \pi_k = 1$ and each component distribution f_k is a product distribution over $X_1, ..., X_p$ parameterized by parameters $\theta_k = (\theta_{k1}, ..., \theta_{kp})$

$$f_k(x_i|\theta_k) = \prod_{j=1}^{p} P_j(x_{ij}|\theta_{kj}). \tag{2}$$

The complete parameterization of the mixture is then given by $\theta = (\pi, \theta_1, ..., \theta_k)$. For a data set D of N samples the likelihood under mixture M is given by

$$P(D|M) = \prod_{i=1}^{N} P(x_i). \tag{3}$$

The way the mixture arises from a given MSA is visualized in Fig. 1; 1a) shows an example MSA with four positions and three subgroups $C_1 - C_3$ within the sequences. An abstract representation of the corresponding mixture model is shown in 1b). Here each position of the alignment is modelled by a discrete random variable $X_1 - X_4$ and each cell in the matrix represents a uniquely parameterized discrete distribution with parameters estimated from the amino acids of the sequences assigned to the subgroup at the respective positions.

The central quantity for both the parameter estimation with *Expectation Maximization* (EM) [6] as well as the subgroup assignment is the posterior of component membership given by

$$\tau_{ik} = \frac{\pi_k \; f_k(x_i|\theta_k)}{\sum_{k=1}^{K} \pi_k \; f_k(x_i|\theta_k)}, \tag{4}$$

i.e. τ_{ik} is the probability that a sample x_i was generated by component k.

a)

	1	2	3	4
	L	E	K	-
C_1	D	A	K	D
	D	A	S	D
	A	L	T	G
C_2	A	L	A	G
	A	L	T	D
	I	P	E	T
C_3	I	P	-	T

b)

	X_1	X_2	X_3	X_4
C_1				
C_2				
C_3				

c)

	X_1	X_2	X_3	X_4
C_1				
C_2				
C_3				

Fig. 1. a) Example input MSA. Eight sequences with four positions each divided into three subgroups. b) Model structure matrices for conventional mixture model and c) structure matrix for the CSI mixture model.

The basic idea of the CSI extension to the mixture framework is to automatically adapt model complexity to match the variability observed in the data. This is visualized in Fig. 1. In 1b) the structure matrix for a conventional mixture model is depicted. Each cell represents a uniquely parameterized distribution for each component and sequence position. In opposition to that a CSI model (Fig. 1 c) may assign the same distribution for a position to several components as indicated by the cell spanning multiple rows in the structure matrix. In example C_1 and C_2 share a distribution parameters for position X_1. For position X_2 all components have the same distribution and for position X_4 all components except C_1 have the same parameters. This not only yields a reduced model complexity, it also allows the direct characterization of protein subgroups by the model structure matrix. For instance it can be seen that position X_4 is uniquely characterizing component C_1. For a protein family data set this might indicate that position X_4 is a candidate for functional residue with respect to subgroup C_1.

Formally the CSI mixture model is defined as follows: Given the set $\mathcal{C} = \{1, .., K\}$ of component indexes and sequence positions $X_1, ..., X_p$, let $G = \{g_j\}_{(j=1,...,p)}$ be the CSI structure of the model M. Then $g_j = (g_{j1}, ...g_{jZ_j})$ such that Z_j is the number of parameter subgroups for X_j and each $g_{jr}, r = 1, ..., Z_j$ is a subset of component indexes from $\mathcal{C}$. Thus, each g_j is a partition of $\mathcal{C}$ into disjunct subsets such that each g_{jr} represents a subgroup of components with the same distribution for X_j. The CSI mixture distribution is then obtained by replacing $f_{kj}(x_{ij}|\theta_{kj})$ with $f_{kj}(x_{ij}|\theta_{g_j(k)j})$ in (2) where $g_j(k) = r$ such that $k \in g_{jr}$. Accordingly $\theta_M = (\pi, \theta_{X_1|g_{1r}}, ..., \theta_{X_p|g_{pr}})$ is the model parameterization. $\theta_{X_j|g_{jr}}$ denotes the different parameter sets in the structure for position j. The complete CSI model M is then given by $M = (G, \theta_M)$. Note that we have covered the structure learning algorithm in more detail in a previous publication [9].

2.2 Dirichlet Mixture Priors

In the Bayesian setting the fit of different models to the data is assessed by the model posterior $P(M|D)$ given by

$$P(M|D) \propto P(M)P(D|M),$$

where $P(D|M)$ is the likelihood of the data under M and $P(M)$ is the model prior. For $P(M) = P(K)P(G)$ a simple factored form was used with $P(K) = \gamma^K$ and $P(G) = \prod_{j=1}^{p} \alpha^{Z_j}$. $\gamma < 1$ and $\alpha < 1$ are hyperparameters which determine the strength of the bias for a less complex model introduced by the prior. The likelihood term $P(D|M)$ is given by

$$P(D|M) = P(D|\overrightarrow{\theta}_M)P(\overrightarrow{\theta}_M).$$

Here $P(D|\overrightarrow{\theta}_M)$ is simply the mixture likelihood (1) evaluated at the *maximum a posteriori* (MAP) parameters $\overrightarrow{\theta}_M$ and $P(\overrightarrow{\theta}_M)$ is a conjugate prior over the model parameters.

One choice of $P(\overrightarrow{\theta}_M)$ for discrete distributions θ is a mixture of Dirichlet distributions. A Dirichlet mixture prior (DMP) over a discrete distribution $\theta = (\theta_1, ..., \theta_Q)$ is given by

$$P(\theta) = \sum_{g=1}^{G} q_g \; D_g(\theta|\alpha_g), \tag{5}$$

where D_g is the Dirichlet density parameterized by $\alpha_g = (\alpha_{g1}, ..., \alpha_{gQ}), \alpha_{gz} > 0$,

$$D_g(\theta) = \frac{\Gamma(\sum_{z=1}^{Q} \alpha_{gz})}{\prod_{z=1}^{Q} \Gamma(\alpha_{gz})} \prod_{z=1}^{Q} \theta_z^{\alpha_{gz}-1}.$$

The DMP has a number of attractive properties for the modeling of protein families. Not only does the DMP retain conjugacy to the discrete distribution which guarantees closed form solutions for the parameter estimates, it also allows for a great degree of flexibility in the induced density over the parameter space. This allows for the integration of amino acid similarities in the structure learning procedure.

2.3 Parameter Estimation

As the Dirichlet distribution is conjugate to the multinomial distribution, the MAP estimates for θ can be computed conveniently.

To obtain the MAP for Dirichlet Mixture priors in case of a mixture of discrete distributions we extend the update rules in [23] where the formulas for the single distribution case have been derived in detail. The MAP solution for the distribution over position j in component k, $\theta_{kj} = (\theta_{kj1}, ..., \theta_{kjQ})$, where Q is the size of the alphabet Σ (21 for amino acids plus gap symbol) is given by

$$\theta_{kjz} = \sum_{g=1}^{G} q_g \frac{T_{kjz} + \alpha_{gz}}{T_k + |\alpha_g|} \tag{6}$$

where the $T_{kj} = (T_{kj1}, ..., T_{kjQ})$ are the expected sufficient statistics of mixture component k in feature j with

$$T_{kjz} = \sum_{i=1}^{N} \tau_{ki} \; \delta_{(x_{ij}=\Sigma_z)},$$

$T_{kj} = \sum_{z=1}^{Q} T_{kjz}$ and q_{gj} is the component membership posterior of θ_{kjz} under the DMP $P(\theta)$ computed according to (4).

2.4 Prior Parameter Derivation

In order to apply the DMP framework on the problem of regularizing the structure learning for protein families we have to specify the parameterization of $P(\theta)$. This includes the choice of G, the q_g and the α_g.

We considered three different approaches to arrive at choices for these parameters,

1. choice of parameters based on a PAM series amino acid substitution probability matrix,
2. use of previously published DMP regularizers [23] based on machine learning techniques and
3. heuristic parameter derivation based on basic chemical properties of the amino acids.

The latter approach proved to be most suitable for our purposes and therefore will be described in more detail below. It should be stressed however that the non-optimal performance of the DMPs from [23] may be caused by their focus on providing suitable regularization to compensate for small sample sizes. While this is certainly related, it is not quite the same as the kind of regularization we require for the CSI structure learning. Clearly a machine learning approach for specifying the prior parameters would be desirable. This however is not straightforward for two reasons: First, it is not clear how the training data for learning a DMP for this application would have to be assembled and secondly the optimization of DMPs is a difficult problem as many local minima exist [23]. In any case, it seems appealing to use a simple heuristically specified DMP in this first analysis in order to establish a baseline performance of the CSI mixtures in this application.

Table 1. The twenty amino acids can be characterized by nine chemical properties. A x in the table denotes the presence, a · the absence of a trait.

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

The impact of an amino acid substitution on the fold of a protein depends on the similarity of the chemical properties of the two amino acids. The more dissimilar the amino acids are, the more pronounced the effect on protein structure will be. The relevant chemical properties can be arranged into a hierarchy of more general and specific properties [18]. The nine properties we consider and the assignment of amino acids is summarized in Table 1. Here 'x' and '·' denote presence and absence of a property respectively. Note that the gap symbol '-' is negative for all properties.

Based on this characterization of the amino acids by their basic chemical properties we construct a DMP as follows: To each of the properties in Table 1 we assign a component D_g in the DMP. The parameters α_g are chosen such that α_{gj}

is larger if amino acid j has the property. This means we construct nine Dirichlet distributions which give high density to $\theta_{X_j|g_{jr}}$ with strong prevalence of amino acids with a certain property. The combination of all property specific D_g in the DMP then yields a density which allows the quantification of similarity between amino acids in the probabilistic framework. In order to arrive at a scheme to choose the parameters of the DMP the following constraints were taken into consideration:

- The strength of a Dirichlet distribution prior D_g is determined by the sum of its parameters $|\alpha_g|$. The size of $|\alpha_g|$ is also anti-proportional to the variance of D_g. To assign equal strength to all property specific Dirichlets D_g, all $|\alpha_g|$ are set to be identical.
- More general properties should receive greater weights q_g in the DMP.
- The strength of the prior, i.e. $|\alpha_g|$ should depend on the size of the data set N.

This leads to the following heuristics for choosing the DMP parameters: Let the strength of each D_g be one tenth of the data set size; i.e. $|\alpha_g| = \frac{N}{10}$ and $b = \frac{0.75\ |\alpha_g|}{21}$ the base value for the parameters α_g. Then $\alpha_{gj} = b$, for all amino acids were the property is absent and

$$\alpha_{gj} = b + \frac{0.25\ |\alpha_g|}{B_g},$$

for all amino acids where the property is present, where B_g denotes the number amino acid which have the property. Finally, the weights q_g are set to

$$q_g = \frac{B_g}{\sum_{g=1}^{G} B_g}$$

which means that more general properties receive proportionally higher weight in the prior. Thus, the priors in the model introduce two types of bias' into the structure learning. An unspecific preference for a less complex model given by $P(M)$ and a specific preference for parameters $\theta_{X_j|g_{jr}}$ that match the amino acid properties encoded in the prior $P(\theta)$.

2.5 Feature Ranking

To predict which features are functional residues for a given subgroup, it is necessary to refine the information in the CSI structure matrix by ranking the informative features. Since these features are distinguished by subgroup specific sequence conservation, the relative entropy is a natural choice to score for putative functional residues.

In order to quantify the relevance of X_j for subgroup i we assume a CSI structure in which X_j is uniquely discriminative for component i, i.e. $Z_j = 2$ with $g_{j1} = \{i\}$ and $g_{j2} = \{1, ..., K\} \setminus i$. Based on this structure a component-specific parameter set θ_{ji} and a parameter set for all other components θ_{other} are constructed by doing a single structural EM update.

The score for feature X_j in component i is then given by $S_{ij} = KL(\theta_{ji}, \theta_{other})$, where KL is the symmetric relative entropy. Note that this is somewhat similar to the setup used in [12]. The major difference being that in [12] subgroup assignments were assumed to be known and in this work the scoring is based on the posterior distribution of component membership and parameter estimates induced by the expected sufficient statistics in the structural EM framework.

3 Results

We evaluated the performance of CSI mixture models for protein subfamilies on a number of data set of different sizes from families with known subtype assignments and structural information. This allows for a validation of the clustering results. Any column in the alignment with more than 33% gaps was removed prior to the clustering. Model selection was carried out using the *Normalized entropy criterion* (NEC) [2]. To assess the impact of the DMP on model performance sensitivity and specificity of the clusterings with DMP were compared to mixtures with the same number of components but a simple uninformative single Dirichlet prior.

3.1 L-Lactate Dehydrogenase Family

We analyzed members of the L-lactate dehydrogenase family, which differ in their substrate specificity. We analyzed two subfamilies, malate and lactate dehydrogenases. In this family, despite substantial variance within the subfamilies and between them, a single position is responsible for defining substrate specificity. Taking PDB 1IB6 as reference sequence, an R in position 81 confers specificity for lactate whereas a Q in the same position would change the substrate to malate. Clusterings were computed for the 29 sequences in the PFAM seed alignment of that domain (PF00056). The alignment contained 16 lactate dehydrogenases (LDH) and 13 malate dehydrogenases (MDH). NEC model selection indicated 2 components to provide the best fit for the data. The two components separated the MDH/LDH groups without error for the DMP mixture.When using the uninformative prior, considerably lower sensitivities and specificities of around 75% were achieved. To assess the robustness of this result we repeatedly trained two component models with DMP and uninformative priors. Averaged over 10 models the DMP achieved sensitivity 95% (SD 1.8) and specificity 93% (SD 2.4), the uninformative prior yielded sensitivity 76% (SD 8.7) and specificity 75% (SD 9.3). Thus, our method was able to identify the two subfamilies correctly without any prior biological knowledge. The position identified as most informative for distinguishing the groups was indeed the one responsible for substrate specificity. Many of the other highly ranked residues were arranged around the NAD interaction site of the domain, which suggests they play a role in malate / lactate recognition.

3.2 Protein Kinase Family

The protein kinase super family is one of the largest and best studied protein families. The human genome contains more than 500 protein kinases [19] with many involved in different diseases like cancer or diabetes. The probably most prominent classification of this key players in signal transduction is between tyrosine and serine/threonine kinases. These can be further subdivided according to different regulatory mechanisms [13]. In our test case, we combined these levels of classifications by joining tyrosine kinases (TK) with two groups of serine threonine kinases, STE (Homologs of yeast Sterile 7, Sterile 11, Sterile 20 kinases) and AGC (Containing PKA, PKG and PKC families). An alignment of 1221 representative sequences of the subfamilies was obtained from the *Protein kinase resource* [24]. The three best NEC model selection scores were assigned to 2, 3 and 4 components. Since the scores were too similar for a clear choice of components, we will consider the clustering of all three models in the following. In the three component model each family acquired its own subgroup with a sensitivity of 79% and a specificity of 83%. Results for the uninformative prior were only slightly worse (about 1% in both sensitivity and specificity) for this data set. These results were highly robust in the repetitions with standard deviations of 0.1%-0.6% on the sensitivities and specificities of both prior types. In the following PDB 2cpk (cAMP-dependent protein kinase, alpha-catalytic subunit, Mus musculus) is used as reference sequence for residue numbering. A ranking of the informative features of the three component model with respect to the TK subgroup yielded within the top 20 positions a region of three residues (168-170) which has been experimentally shown to be important for kinase substrate specificity [11]. For the two component model the TK and STE sequences were collected in one subgroup and the second was almost exclusively AGC. The four component model finally yielded a high specificity clustering (98%) in which the AGC sequence got split over two components. The sensitivity was 76%.

3.3 Nucleotidyl Cyclase Family

Nucleotidyl cyclases play an important role in cellular signaling by producing the second messengers cAMP and cGMP which regulate the activity of many other signalling molecules. As cGMP and cAMP fulfill different biological roles, specificity of converting enzymes is imperative. Five residues have been experimentally confirmed to convey substrate specificity, namely 938, 1016, 1018, 1019, 1020 (numbering according to PDB 1AB8) [17]. We used this family as a test case for families with multiple sites involved in functional classification, complementing the L-lactate dehydrogenase family with a single site. We computed a MSA from 132 GC (EC 4.6.1.2) and AC (EC 4.6.1.1) sequences obtained from the ExPASy data base [8]. The NEC model selection indicated two components to provide the best fit. The model with optimal NEC produced a clustering with sensitivity of 83% and specificity 87% with respect to the GC / AC subgroups. For the uninformative prior these values decreased to 70% and 73% respectively. Averaged over 10 models the uninformative prior yielded a decreased performance of 59% (SD 5.3) sensitivity and 62% (SD 5.6) respectively. The averaged

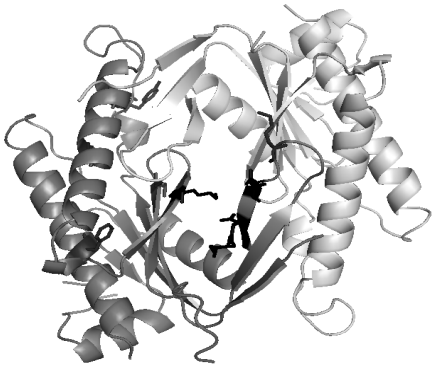

Fig. 2. Adenylyl cyclase with classifying sites highlighted - Subunit I in dark grey, subunit II in light grey. The 10 most informative sites were selected. Shown in black: experimentally validated identified sites, darkest grey: additional identified sites. A colored version of the figure is available from *http://algorithmics.molgen.mpg.de/pymix/Figure_Cyclase.png*

results for the DMP were sensitivity 73% (SD 4.3) and specificity 77% (SD 4.8). Figure 2 shows the three dimensional structure of 1AB8 with the 10 most informative sites highlighted. Indeed, these contain 4 of the sites involved in substrate specificity (1018 (ranked 2.), 1016 (3.), 938 (6.), 1019 (9.)). Further top ranking positions included sites which are part of the subunit I and II domain interface (919, 912, 911). Position 943 is right next to a forsoklin interaction site and position 891 interacts with magnesium. Residue 921 finally, is also a metal interacting site [27]. Thus, not only known substrate specific sites were identified, but also further functional sites. It would be interesting to experimentally test identified sites with no functional annotation.

4 Discussion

The results of CSI mixture-based clustering on a number of different protein families show that the approach is capable of simultaneously finding biological relevant subgroups, as well as predicting functional residues that characterize these groups. The functional residue prediction proved to be robust to some degree to imperfections in the clustering. This implies that our unsupervised approach to simultaneous clustering and determination of functional residues is feasible. Also note that our results for the functional residue prediction are strongly consistent with those reported by studies using supervised methods on the same families [12,26]. With regard to experimentally confirmed specificity determining residues found by these studies, we found 1/1 for L-lactate dehydrogenase, 3/3 for protein kinases and 4/5 for nucleotidyl cyclase.

The results also show that the DMP used in this analysis, in spite of being based on basic chemical properties and simple heuristics, consistently increases the performance of the mixture framework for the application on protein data,

although the degree of improvement differs considerably between the families. This is not unexpected as one would expect differing amounts of synonymous substitutions within the various subgroups and that is the situation where the DMP makes the largest difference as compared to the uninformative prior. For comparison we also applied the tree-based method SPEL [22] to our data. The sources were obtained from the authors and run with default parameters. For the MDH/LDH data the true functional position 81 was not among the ten positions returned by SPEL. For the two larger data sets, there were implementation-issues and SPEL did unfortunately not produce any results.

For future work it might be worth investigating the impact of different DMPs on the clustering results and in particular whether customized DMPs for specific applications yield improvement over the more general purpose prior used in this work. Moreover, now that the usefulness of the method has been established on families with abundant prior knowledge about subgroups and structure, the next step must be to bring the method to bear to predict groups and functional residues on data sets where such knowledge does not exist yet. Finally, the software we developed to carry out this analysis *PyMix - the Python Mixture Package* is available from our home page *http://algorithmics.molgen.mpg.de/pymix.html*.

References

1. Barash, Y., Friedman, N.: Context-specific bayesian clustering for gene expression data. J. Comput. Biol. 9(2), 169–191 (2002)
2. Biernacki, C., Celeux, G., Govaert, G.: An improvement of the NEC criterion for assessing the number of clusters in a mixture model. Non-Linear Anal. 20(3), 267–272 (1999)
3. Boutilier, C., Friedman, N., Goldszmidt, M., Koller, D.: Context-specific independence in Bayesian networks. In: Uncertainty in Artificial Intelligence, pp. 115–123 (1996)
4. Chakrabarti, S., Lanczycki, C.J.: Analysis and prediction of functionally important sites in proteins. Protein Sci. 16(1), 4–13 (2007)
5. Chickering, D.M., Heckerman, D.: Efficient approximations for the marginal likelihood of bayesian networks with hidden variables. Mach. Learn. 29(2-3), 181–212 (1997)
6. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 1–38 (1977)
7. Friedman, N., Goldszmidt, M.: Learning bayesian networks with local structure. In: Proceedings of the NATO Advanced Study Institute on Learning in graphical models, pp. 421–459. Kluwer Academic Publishers, Norwell, MA, USA (1998)
8. Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, D., Bairoch, A.: ExPASy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res. 31(13), 3784–3788 (2003)
9. Georgi, B., Schliep, A.: Context-specific independence mixture modeling for positional weight matrices. Bioinformatics 22(14), e166–173 (2006)
10. Georgi, B., Spence, M.A., Flodman, P., Schliep, A.: Mixture model based group inference in fused genotype and phenotype data. In: Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Heidelberg (2007)

11. Hanks, S.K., Quinn, A.M., Hunter, T.: The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. Science 241(4861), 42–52 (1988)
12. Hannenhalli, S., Russell, R.B.: Analysis and prediction of functional sub-types from protein sequence alignments. J. Mol. Biol. 303(1), 61–76 (2000)
13. Hunter, T.: Protein kinase classification. Methods Enzymol 200, 3–37 (1991)
14. Jones, S., Thornton, J.M.: Searching for functional sites in protein structures. Curr. Opin. Chem. Biol. 8(1), 3–7 (2004)
15. Lazareva-Ulitsky, B., Diemer, K., Thomas, P.D.: On the quality of tree-based protein classification. Bioinformatics 21(9), 1876–1890 (2005) Comparative Study
16. Lichtarge, O., Bourne, H.R., Cohen, F.E.: An evolutionary trace method defines binding surfaces common to protein families. J. Mol. Biol. 257(2), 342–358 (1996)
17. Liu, Y., Ruoho, A.E., Rao, V.D., Hurley, J.H.: Catalytic mechanism of the adenylyl and guanylyl cyclases: modeling and mutational analysis. Proc. Natl. Acad. Sci. USA 94(25), 13414–13419 (1997)
18. Livingstone, C.D., Barton, G.J.: Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. Comput. Appl. Biosci. 9(6), 745–756 (1993)
19. Manning, G., Whyte, D.B., Martinez, R., Hunter, T., Sudarsanam, S.: The protein kinase complement of the human genome. Science 298(5600), 1912–1934 (2002)
20. McLachlan, G.J., Peel, D.: Finite Mixture Models. John Wiley & Sons, Chichester (2000)
21. Mirny, L.A., Gelfand, M.S.: Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. J. Mol. Biol. 321(1), 7–20 (2002)
22. Pei, J., Cai, W., Kinch, L.N, Grishin, N.V.: Prediction of functional specificity determinants from protein sequences using log-likelihood ratios. Bioinformatics 22(2), 164–171 (2006)
23. Sjolander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I.S, Haussler, D.: Dirichlet mixtures: A method for improving detection of weak but significant protein sequence homology. Technical report, University of California at Santa Cruz, Santa Cruz, CA, USA (1996)
24. Smith, C.M., Shindyalov, I.N., Veretnik, S., Gribskov, M., Taylor, S.S., Ten Eyck, L.F., Bourne, P.E.: The protein kinase resource. Trends Biochem. Sci. 22(11), 444–446 (1997)
25. Wicker, N., Perrin, G.R., Thierry, J.C., Poch, O.: Secator: a program for inferring protein subfamilies from phylogenetic trees. Mol. Biol. Evol. 18(8), 1435–1441 (2001)
26. Yu, G., Park, B.-H., Chandramohan, P., Munavalli, R., Geist, A., Samatova, N.F.: In silico discovery of enzyme-substrate specificity-determining residue clusters. J. Mol. Biol. 352(5), 1105–1117 (2005)
27. Zhang, G., Liu, Y., Ruoho, A.E., Hurley, J.H.: Structure of the adenylyl cyclase catalytic core. Nature 386(6622), 247–253 (1997)