Keilwagen et al. BMC Bioinformatics 2010, 11:98
http://www.biomedcentral.com/1471-2105/11/98

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Unifying generative and discriminative learning principles

Jens Keilwagen[1*], Jan Grau[2], Stefan Posch[2], Marc Strickert[1], Ivo Grosse[2]

Abstract

Background: The recognition of functional binding sites in genomic DNA remains one of the fundamental challenges of genome research. During the last decades, a plethora of different and well-adapted models has been developed, but only little attention has been payed to the development of different and similarly well-adapted learning principles. Only recently it was noticed that discriminative learning principles can be superior over generative ones in diverse bioinformatics applications, too.

Results: Here, we propose a generalization of generative and discriminative learning principles containing the maximum likelihood, maximum a posteriori, maximum conditional likelihood, maximum supervised posterior, generative-discriminative trade-off, and penalized generative-discriminative trade-off learning principles as special cases, and we illustrate its efficacy for the recognition of vertebrate transcription factor binding sites.

Conclusions: We find that the proposed learning principle helps to improve the recognition of transcription factor binding sites, enabling better computational approaches for extracting as much information as possible from valuable wet-lab data. We make all implementations available in the open-source library Jstacs so that this learning principle can be easily applied to other classification problems in the field of genome and epigenome analysis.

Background

Classification of unlabeled data is one of the main tasks in bioinformatics. For DNA sequence analysis, this classification task is synonymous to the computational recognition of short signal sequences in genomic DNA. Examples include the recognition of transcription factor binding sites (TFBSs) [1,2], transcription start sites [3,4], donor or acceptor splice sites [5-7], nucleosome binding sites [8,9], miRNA binding sites [10,11], or binding sites of insulators like CTCF [12].

Many of the employed algorithms use statistical models for representing the distribution of sequences. These models range from simple models like the position weight matrix (PWM) model [1,13,14], the weight array matrix (WAM) model [6,8,15], or Markov models of higher order [16,17] to complex models like Bayesian networks [2,18,19] or Markov random fields [7,20,21]. A wealth of different models has been proposed for different data sets and different biological questions, and it is advisable to carefully choose an appropriate model for each data set and each biological question separately [4,7,22]. However, the performance of a model highly depends on the model parameters learned from training data. In comparison to the effort spent for developing and choosing appropriate models, developing and choosing appropriate learning principles has been neglected, even though this choice is of fundamental importance [23-27] and equally non-trivial.

In the last decades, several learning principles have been proposed for estimating model parameters. The *maximum likelihood* (ML) learning principle [28,29] is one of the first and most popular learning principles used in bioinformatics. An alternative is the *maximum a posteriori* (MAP) learning principle [30] that applies a prior density to the parameters of the models.

The ML and the MAP learning principles are commonly referred to as *generative*. Recently, *discriminative* learning principles have been shown to be promising in several bioinformatics applications [16,17,20,26,27,31]. The discriminative analogue to the ML learning principle is the *maximum conditional likelihood* (MCL) learning principle [24,25,32-34], and the *maximum supervised posterior* (MSP) learning principle [35,36] has

* Correspondence: Jens.Keilwagen@ipk-gatersleben.de
[1]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany

been proposed as discriminative analogue to the MAP learning principle.

In addition to these four learning principles, hybrid learning principles have been proposed to combine the advantages of generative and discriminative learning principles [37-42]. Specifically, the *generative-discriminative trade-off* (GDT) learning principle that interpolates between the ML and the MCL learning principle has been proposed in [37], and the *penalized generative-discriminative trade-off* (PGDT) learning principle that interpolates between the MAP and the MSP learning principle has been proposed in [41].

Here, we introduce a unified generative-discriminative learning principle containing the ML, the MAP, the MCL, the MSP, the GDT, and the PGDT learning principle as limiting cases. We discuss the interpretation of this learning principle, and we investigate its utility using four data sets of TFBSs.

Results and Discussion

In this section, we present six established learning principles, then introduce the unified generative-discriminative learning principle containing the six established learning principles as special cases, and finally present some discussion and interpretation of the learning principle introduced. We start with considering classifiers that are based on probabilistic models defined by the likelihood $P(\underline{x} \mid c, \underline{\lambda})$ for sequence $\underline{x}$ given class label c and parameter vector $\underline{\lambda}$. Based on such models the decision criterion [43] of the classifier is defined as

$$\begin{aligned}\hat{c} &:= \arg\max_{c\in\mathcal{C}} P(c \mid \underline{x}, \underline{\lambda}) = \arg\max_{c\in\mathcal{C}} P(c, \underline{x} \mid \underline{\lambda}) \\ &= \arg\max_{c\in\mathcal{C}} P(c \mid \underline{\lambda}) \cdot P(\underline{x} \mid c, \underline{\lambda}),\end{aligned} \tag{1}$$

where $P(c \mid \underline{x}, \underline{\lambda})$ is the conditional likelihood of class label c given sequence $\underline{x}$ and parameter vector $\underline{\lambda}$, $P(c, \underline{x} \mid \underline{\lambda})$ is the likelihood of sequence $\underline{x}$ and class label c given parameter vector $\underline{\lambda}$, $P(c \mid \underline{\lambda})$ is the probability of class c given parameter vector $\underline{\lambda}$, and $P(\underline{x} \mid c, \underline{\lambda})$ is the conditional probability of sequence $\underline{x}$ given class label c and parameter vector $\underline{\lambda}$.

The decision and classification performance depend on the parameter vector $\underline{\lambda}$. Hence, one needs to infer *appropriate* parameter vectors $\underline{\lambda}$ from a data set $\underline{D} := (\underline{x}_1, \ldots, \underline{x}_N)$ of N statistically independent and identically distributed (i.i.d.) sequences and the corresponding class labels $\underline{C} := (c_1, \ldots, c_N)$. In the first subsection, we present six learning principles that have been proposed in the machine-learning community and that are nowadays also used in bioinformatics. In the second subsection, we propose a unified learning principle containing all of these six learning principles as special cases. In the third subsection, we provide a mathematical interpretation of this learning principle, and in the fourth subsection we present four case studies illustrating the utility of this learning principle. We present some implementation details in the **Methods** section.

Established learning principles

Learning principles can be categorized by two criteria. On the one hand, they can be divided by their objective into generative, discriminative, and hybrid learning principles. Generative learning principles aim at an accurate representation of the distribution of the training data in each of the classes, discriminative learning principles aim at an accurate classification of the training data into the classes, and hybrid learning principles are an interpolation between generative and discriminative learning principles. On the other hand, learning principles can be divided by their utilization of prior knowledge into Bayesian and non Bayesian. We call learning principles that incorporate a prior density $Q(\underline{\lambda} \mid \underline{\alpha})$ on the parameter vector $\underline{\lambda}$ Bayesian, where $\underline{\alpha}$ denotes a vector of hyper parameters, while we call learning principles that only use the data - without any prior - to estimate the parameter vector non Bayesian. In Table 1 we present six established learning principles and their categorization by the above-mentioned criteria, and we describe these learning principles in more detail in the remainder of this subsection.

Generative learning principles

The maximum likelihood (ML) learning principle is one of the first learning principles used in bioinformatics. Originally, it was proposed by R. A. Fisher at the beginning of the 20^{th} century [28,29]. The ML learning principle aims at finding the parameter vector $\hat{\underline{\lambda}}_{\mathrm{ML}}$ that maximizes the likelihood of the labeled data set $(\underline{C}, \underline{D})$ given the parameter vector $\underline{\lambda}$,

$$\hat{\underline{\lambda}}_{\mathrm{ML}} := \arg\max_{\underline{\lambda}} P(\underline{C}, \underline{D} \mid \underline{\lambda}). \tag{2}$$

Table 1 Learning principles

		prior knowledge	
		non Bayesian	Bayesian
objective	generative	ML	MAP
	hybrid	GDT	PGDT
	discriminative	MCL	MSP

The table shows six established learning principles that can be grouped by their objective as being generative, hybrid, or discriminative and utilization of prior knowledge with the two possibilities non Bayesian and Bayesian. The four elementary learning principles are the generative, non Bayesian maximum likelihood (ML) learning principle, the generative, Bayesian maximum a posteriori (MAP) learning principle, the discriminative, non Bayesian maximum conditional likelihood (MCL) learning principle, and the discriminative, Bayesian maximum supervised posterior (MSP) learning principle. The hybrid learning principles which interpolate between generative and discriminative learning principles are the non Bayesian generative-discriminative trade-off (GDT) learning principle and the penalized generative-discriminative trade-off (PGDT) learning principle.

However, for many applications, the amount of sequence data available for training is very limited. For this reason, the ML learning principle often leads to suboptimal classification performance e.g. due to zero-occurrences of some nucleotides or oligonucleotides in the training data sets.

The maximum a posteriori (MAP) learning principle, which applies a prior $Q(\underline{\lambda}|\underline{\alpha})$ to the parameter vector, establishes a theoretical foundation to alleviate this problem and at the same time allows the inclusion of prior knowledge aside from the training data [30]. The MAP learning principle aims at finding the parameter vector $\hat{\underline{\lambda}}_{\mathrm{MAP}}$ that maximizes the posterior density,

$$\hat{\underline{\lambda}}_{\mathrm{MAP}} := \arg\max_{\underline{\lambda}} P(\underline{\lambda} \mid \underline{C}, \underline{D}, \underline{\alpha}\,|) = \arg\max_{\underline{\lambda}} P(\underline{C}, \underline{D} \mid \underline{\lambda}\,|) \cdot Q(\underline{\lambda} \mid \underline{\alpha}). \quad (3)$$

If for a given family of likelihood functions $P(\underline{C}, \underline{D}|\underline{\lambda})$ the posterior $P(\underline{\lambda}|\underline{C}, \underline{D}, \underline{\alpha})$ is in the same family of distributions as the prior $Q(\underline{\lambda}|\underline{\alpha})$, i.e., if

$$Q(\underline{\lambda} \mid \tilde{\underline{\alpha}}) = P(\underline{\lambda} \mid \underline{C}, \underline{D}, \underline{\alpha}) \propto P(\underline{C}, \underline{D} \mid \underline{\lambda}) \cdot Q(\underline{\lambda} \mid \underline{\alpha}) \quad (4)$$

the prior is said to be *conjugate* to this family of likelihood functions, for hyper parameter vector $\tilde{\underline{\alpha}}$ incorporates both prior knowledge and training data. Conjugate priors often allow an interpretation of the hyper parameter vector as stemming from an a priorily observed set of *"pseudo data."* In addition, it allows finding the optimal parameter vector $\hat{\underline{\lambda}}_{\mathrm{MAP}}$ analytically provided one can determine the maximum of the prior analytically.

Discriminative learning principles

Discriminative learning principles have been shown to be promising in the field of bioinformatics [16,17,20,26,31]. The discriminative analogue to the ML learning principle is the maximum conditional likelihood (MCL) learning principle [24,25,32-34] that aims at finding the parameter vector $\hat{\underline{\lambda}}_{\mathrm{MCL}}$ that maximizes the conditional likelihood of the labels $\underline{C}$ given the data $\underline{D}$ and parameter vector $\underline{\lambda}$,

$$\hat{\underline{\lambda}}_{\mathrm{MCL}} := \arg\max_{\underline{\lambda}} P(\underline{C} \mid \underline{D}, \underline{\lambda}). \quad (5)$$

The effects of limited data may be even more severe when using the MCL learning principle compared to generative learning principles [23]. To overcome this problem, the maximum supervised posterior (MSP) learning principle [35,36] has been proposed as discriminative analogue to the MAP learning principle. In analogy to equation (3), the MSP learning principle aims at finding the parameter vector $\hat{\underline{\lambda}}_{\mathrm{MSP}}$ that maximizes the product of the conditional likelihood and the prior density,

$$\hat{\underline{\lambda}}_{\mathrm{MSP}} := \arg\max_{\underline{\lambda}} P(\underline{C} \mid \underline{D}, \underline{\lambda}) \cdot Q(\underline{\lambda} \mid \underline{\alpha}). \quad (6)$$

Generative-discriminative trade-offs

Different hybrid learning principles have been proposed in the machine learning community [37,39,41]. Hybrid learning principles aim at combining the strengths of generative and discriminative learning principles. Here, we follow the ideas of Bouchard and co-workers who propose an interpolation between the generative ML learning principle and the discriminative MCL learning principle [37] as well as the generative MAP learning principle and the discriminative MSP learning principle [41]. The generative-discriminative trade-off (GDT) learning principle proposed in [37] aims at finding the parameter vector $\underline{\lambda}$ that maximizes the weighted product of the conditional likelihood and likelihood, i.e.,

$$\hat{\underline{\lambda}}_{\mathrm{GDT}} := \arg\max_{\underline{\lambda}} P(\underline{C} \mid \underline{D}, \underline{\lambda})^{1-\gamma} \cdot P(\underline{C}, \underline{D} \mid \underline{\lambda})^{\gamma} \quad (7)$$

for given weight $\gamma \in [0, 1]$. As special cases of the PGDT learning principle, we obtain the ML learning principle for $\gamma = 1$ and the MCL learning principle for $\gamma = 0$. By varying γ between 0 and 1, different beneficial trade-offs can be obtained for classification.

In close analogy to the MAP and the MSP learning principle, which are obtained by multiplying a prior to the likelihood and conditional likelihood, respectively, the penalized generative-discriminative trade-off (PGDT) learning principle aims at finding the parameter vector $\underline{\lambda}$ that maximizes the objective function

$$\hat{\underline{\lambda}}_{\mathrm{PGDT}} := \arg\max_{\underline{\lambda}} P(\underline{C} \mid \underline{D}, \underline{\lambda})^{1-\gamma} \cdot P(\underline{C}, \underline{D} \mid \underline{\lambda})^{\gamma} \cdot Q(\underline{\lambda} \mid \underline{\alpha}) \quad (8)$$

for given weight $\gamma \in [0, 1]$. As special cases of the PGDT learning principle, we obtain the MAP learning principle for $\gamma = 1$ and the MSP learning principle for $\gamma = 0$.

We summarize the six established learning principles in Table 1.

Unified generative-discriminative learning principle

Comparing equations (2), (3), (5), (6), (7), and (8), we find that the following three terms are sufficient for defining these six learning principles:

1. the conditional likelihood $P(\underline{C}|\underline{D}, \underline{\lambda})$,
2. the likelihood $P(\underline{C}, \underline{D}|\underline{\lambda})$, and
3. the prior $Q(\underline{\lambda}|\underline{\alpha})$.

With the goal of unifying and generalizing all six learning principles, we propose a unified generative-discriminative learning principle that aims at finding the parameter vector $\underline{\lambda}$ that maximizes the weighted product of the conditional likelihood, likelihood, and prior, i.e.,

$$\hat{\underline{\lambda}} := \arg\max_{\underline{\lambda}} P(\underline{C} \mid \underline{D}, \underline{\lambda})^{\beta_0} \cdot P(\underline{C}, \underline{D} \mid \underline{\lambda})^{\beta_1} \cdot Q(\underline{\lambda} \mid \underline{\alpha})^{\beta_2} \quad (9)$$

with the weighting factors $\underline{\beta} := (\beta_0, \beta_1, \beta_2)$, $\beta_0, \beta_1, \beta_2 \in \mathbb{R}_0^+$, and $\beta_0 + \beta_1 + \beta_2 = 1$.

The six established learning principles can be obtained as limiting cases of equation (9) as follows

- ML if $\underline{\beta} = (0, 1, 0)$,
- MAP if $\underline{\beta} = (0, 0.5, 0.5)$,
- MCL if $\underline{\beta} = (1, 0, 0)$,
- MSP if $\underline{\beta} = (0.5, 0, 0.5)$,
- GDT if $\beta_2 = 0$, and
- PGDT if $\beta_2 = 0.5$.

In Figure 1(a), we illustrate the simplex $\underline{\beta}$ by a projection onto the (β_0, β_1)-plane showing the established learning principles as well as the unified generative-discriminative learning principle. However, there are several other hybrid learning principles that are not covered by this unification.

Interpretation of the unified generative-discriminative learning principle

In this subsection, we investigate the simplex $\underline{\beta}$ and its relation to six established learning principles. First, we consider the axes of the simplex $\underline{\beta}$. We can write the learning principle that corresponds to the β_0-axis ($\beta_0 > 0$ and $\beta_1 = 0$) using the constraint $\beta_0 = 1 - \beta_2$ for this axis as

$$\hat{\underline{\lambda}} = \arg\max_{\underline{\lambda}} P(\underline{C} \mid \underline{D}, \underline{\lambda}) \cdot Q(\underline{\lambda} \mid \underline{\alpha})^{\frac{\beta_2}{1-\beta_2}}. \quad (10a)$$

Similarly, we can write the learning principle that corresponds to the β_1-axis (with $\beta_0 = 0$ and $\beta_1 > 0$) as

$$\hat{\underline{\lambda}} = \arg\max_{\underline{\lambda}} P(\underline{C}, \underline{D} \mid \underline{\lambda}) \cdot Q(\underline{\lambda} \mid \underline{\alpha})^{\frac{\beta_2}{1-\beta_2}}. \quad (10b)$$

These equations state that each point on the abscissa (β_0-axis) and on the ordinate (β_1-axis) corresponds to the MSP and the MAP learning principle, respectively, with a weighted prior.

If the prior fulfills the condition

$$Q(\underline{\lambda} \mid \underline{\alpha})^{\xi} \propto Q(\underline{\lambda} \mid \xi\underline{\alpha}) \quad (11)$$

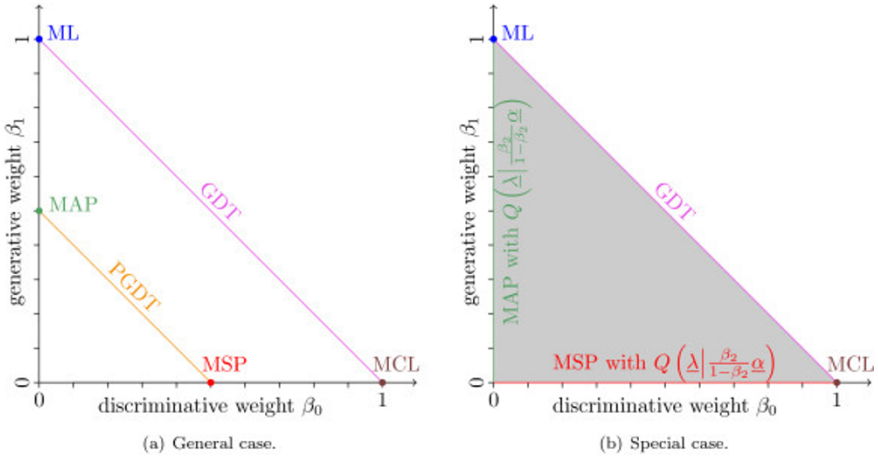

Figure 1 Illustration of the unified generative-discriminative learning principle. The plots show a projection of the simplex $\underline{\beta}$ onto the (β_0, β_1)-plane and the corresponding learning principles for the specific weights encoded by colors. Figure 1(a) shows the general interpretation of the simplex where the points (0, 1), (0, 0.5), (1, 0), and (0.5, 0) refer to the ML, MAP, MCL, and MSP learning principle, respectively, while the lines $\beta_1 = 1 - \beta_0$ and $\beta_1 = 0.5 - \beta_0$ refer to the GDT and PGDT learning principle, respectively. Figure 1(b) shows the interpretation of the unified generative-discriminative learning principle for a conjugate prior that satisfies the condition of equation (11). In this case, each point on the abscissa (β_0-axis) and ordinate (β_1-axis) refers to the MSP and MAP learning principle, respectively, using the prior in a weighted version $\underline{\lambda}\left|\frac{\beta_2}{1-\beta_2}\underline{\alpha}\right.$. The simplex colored in gray corresponds to the MSP learning principle using the weighted posterior $\underline{\lambda}\left|\frac{\beta_1}{\beta_0}\tilde{\underline{\alpha}}\right.$ as prior for the parameter vector $\underline{\lambda}$.

for any $\xi \in \mathbb{R}^+$, each point (1 - β_2, 0) and (0, 1 - β_2) on the axes corresponds to either the MSP or the MAP learning principle using the prior $\underline{\lambda}\left|\frac{\beta_2}{1-\beta_2}\underline{\alpha}\right.$, respectively. The *Generalized Dirichlet prior* for Markov random fields [27], which has been proposed to allow a direct comparison of the MAP and the MSP learning principle, fulfills the condition of equation (11) (Appendix A in Additional File 1).

Second, we consider the lines $\beta_1 = v - \beta_0$ with $v \in [0, 1]$. As visualized in Figure 1(a), the unified generative-discriminative learning principle results in the GDT and the PGDT learning principle for $v = 1$ and $v = 0.5$, respectively. Using $\beta_2 \in (0, 1)$ and the condition of equation (11) with $\xi = \frac{\beta_2}{1-\beta_2}$, we find that equation (9) can be written as

$$\hat{\underline{\lambda}} = \arg\max_{\underline{\lambda}} P(\underline{C} \mid \underline{D}, \underline{\lambda})^{\frac{\beta_0}{1-\beta_2}} \cdot P(\underline{C}, \underline{D} \mid \underline{\lambda})^{\frac{\beta_1}{1-\beta_2}} \cdot Q\left(\underline{\lambda}\left|\frac{\beta_1}{1-\beta_2}\cdot\underline{\alpha}\right.\right). \quad (12)$$

This equation is equivalent to equation (8), stating that - for each β_2 - each point on the line $\beta_1 = (1 - \beta_2) - \beta_0$ corresponds to a specific instance of the PGDT learning principle with prior $\underline{\lambda}\left|\frac{\beta_2}{1-\beta_2}\cdot\underline{\alpha}\right.$. Using this result, the unified generative-discriminative learning principle allows an in-depth analysis of the PGDT learning principle using different priors.

Finally, we consider a second interpretation of the unified generative-discriminative learning principle. The last two terms of the equation (9) consisting of the weighted likelihood and the weighted prior might be interpreted as a weighted posterior. Using the assumption of conjugacy (equation (4)), the condition of equation (11), and $\beta_0, \beta_1, \beta_2 \in \mathbb{R}^+$, we obtain

$$\hat{\underline{\lambda}} = \arg\max_{\underline{\lambda}} P(\underline{C} \mid \underline{D}, \underline{\lambda}) \cdot \left[P(\underline{C}, \underline{D} \mid \underline{\lambda}) \cdot Q\left(\underline{\lambda}\left|\frac{\beta_2}{\beta_1}\underline{\alpha}\right.\right)\right]^{\frac{\beta_1}{\beta_0}} \quad (13a)$$

$$= \arg\max_{\underline{\lambda}} P(\underline{C} \mid \underline{D}, \underline{\lambda}) \cdot Q\left(\underline{\lambda}\left|\frac{\beta_1}{\beta_0}\tilde{\underline{\alpha}}\right.\right) \quad (13b)$$

stating that each point on the simplex can be interpreted as MSP learning principle with an informative prior $\underline{\lambda}\left|\frac{\beta_1}{\beta_0}\tilde{\underline{\alpha}}\right.$ composed of the likelihood and the original prior. Interestingly, the interpretation of each point of the simplex as instance of the MSP learning principle using the weighted posterior as prior remains valid even for priors that do not fulfill the conditions. Figure 1(b) visualizes these results.

Testing

In this subsection, we present four case studies illustrating the utility of the unified generative-discriminative learning principle. In specific practical applications, the choice of appropriate training and test data sets is a highly non-trivial task. Since the final results strongly depend on the chosen data sets, we recommend this choice to be made with great care and in a problem-specific manner. This choice is typically influenced by a-priori knowledge on both the expected binding sites (BSs) and the targeted genome regions. Examples of features that are often considered when choosing appropriate data sets are the GC content of the target region, their association with CpG islands, or their size and proximity to transcription start sites.

Carefully choosing appropriate training and test data sets is of additional advantage if the set of targeted genome regions is not homogeneous, e.g., comprising both GC-rich and GC-poor regions, CpG islands and CpG deserts, TATA-containing and TATA-less promoters, upstream regions with and without BSs of another TF, etc. In this case, one often finds that different learning principles work well for different subgroups, even if the same combination of models is chosen, providing the possibility of choosing subgroup-specific learning principles by choosing different values of $\underline{\beta}$.

These considerations are vital for a successful prediction of TFBSs, but beyond the scope of this paper, so we choose some traditional data sets in the following case study. Specifically, we choose the following four data sets of experimentally verified TFBSs of length L = 16 bp from TRANSFAC [44]. The data set AR/GR/PR contains 104 BSs from three specific steroid hormone receptors from the same class of TFs. The data sets GATA and Thyroid contain 110 and 127 BSs, respectively, of TFs with zinc-coordinating DNA-binding domains. Finally, the data set NF-κB contains 72 BSs of the rapid-acting family of primary TFs NFκB. As background data set we choose the standard background data set of TRANSFAC consisting of 267 second exons of human genes with 68,141 bp in total, which we chunk into sequences of length of at most 100 bp. We build classifiers with the goal of classifying, for each family of TFs separately, a given 16-mer as BS or as subsequence of a background sequence.

We choose a naïve Bayes classifier consisting of two PWM models and the Generalized Dirichlet prior [27] using an *equivalent sample size* (ESS) (Appendix A in Additional File 1) of 4 and 1024 for the foreground and the background class, respectively. We choose the sensitivity for a specificity of 99.9% [19] as performance measure. We present the results for three additional performance measures in Appendix B of Additional File 1. We perform a 1,000-fold stratified hold-out sampling

with 90% of the data for training and 10% of the data for assessing the performance measures for the evaluation of the unified generative-discriminative learning principle.

In Figure 2a, we illustrate the results for BSs of the TFs AR/GR/PR. Considering the ML learning principle located at $(\beta_0, \beta_1) = (0, 1)$ and the MCL learning principle located at $(\beta_0, \beta_1) = (1, 0)$, we find a sensitivity of 54.7% and 55.2%, respectively. Interestingly, the MCL learning principle achieves a higher sensitivity for a given specificity of 99.9% than the ML learning principle for this small data set. Using the Generalized Dirichlet prior with hyper parameters corresponding to uniform pseudo data, the sensitivities can be increased. Considering the MAP learning principle located at $(\beta_0, \beta_1) = (0, 0.5)$ and the MSP learning principle located at $(\beta_0, \beta_1) = (0.5, 0)$, we obtain a sensitivity of 54.9% and 55.6%, respectively. This shows that the MSP learning principle yields an increase of sensitivity of 0.7% compared to the MAP learning principle, consistent with the general observation that discriminatively learned classifiers often outperform their generatively learned counterparts. This increase of sensitivity is achieved using the same prior and the same hyper parameters for both learning principles, but it is possible that the particular choice of the hyper parameters may favour one of the learning principles.

Following equations (10a) and (10b), each point on the β_0- and β_1-axis corresponds to the MSP and the MAP learning principle, respectively, with specific hyper parameters $\underline{\alpha}$. The location on the axis indicates the strength of the prior reflected by the *virtual* ESS (Appendix A in Additional File 1). Next, we investigate for both learning principles the influence of the strength of the prior on the sensitivity.

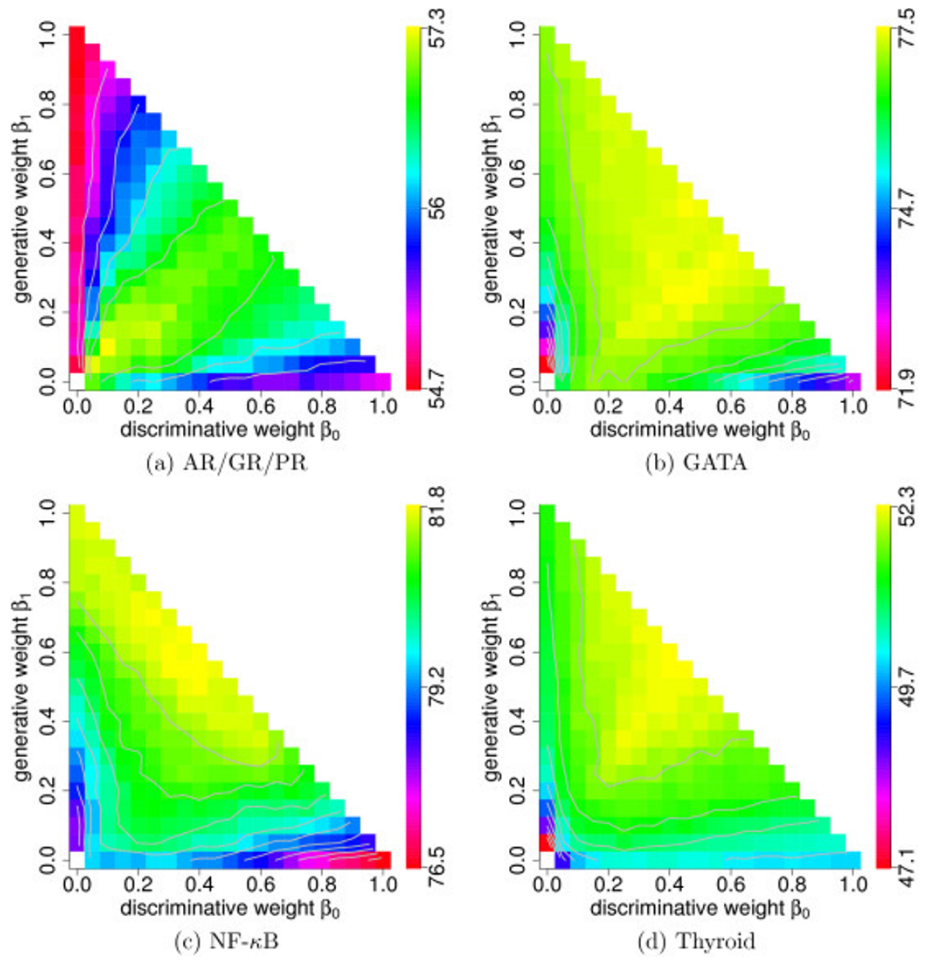

Figure 2 Performance of the unified generative-discriminative learning principle for four data sets. We perform a 1,000-fold stratified hold-out sampling procedure for the four data sets, record for different values of $\underline{\beta}$ the mean sensitivity for a fixed specificity of 99.9%, and plot the mean sensitivities on the simplex $\underline{\beta}$ in analogy to Figure 1. Yellow indicates the highest sensitivity, red indicates the lowest sensitivity, and the gray contour lines of each subfigure indicate multiples of the standard error of the maximum sensitivity.

For the MAP learning principle, the sensitivity ranges from 54.7% for $\underline{\beta}$ = (0, 0.05, 0.95) to 54.8% for $\underline{\beta}$ = (0, 0.95, 0.05), achieving a maximum of 55.1% for $\underline{\beta}$ = (0, 0.1, 0.9). For the MSP learning principle, the sensitivity ranges from the maximum value 56.7% for $\underline{\beta}$ = (0.05, 0, 0.95) to 55.3% for $\underline{\beta}$ = (0.95, 0, 0.05). Comparing the maximum sensitivities for both learning principles and different virtual ESSs, we find that the MSP learning principle with a maximum sensitivity of 56.7% clearly outperforms the MAP learning principle by 1.6%, whereas the difference of sensitivities is only 0.7% for the original ESS.

Investigating this increase in the difference of sensitivities between the results for the MAP and the MSP learning principle, we find that the sensitivity increases for decreasing β_0 on the β_0-axis, which corresponds to the MSP learning principle with an increasing virtual ESS of the prior. In contrast to this observation, the sensitivity for the MAP learning principle increases less strongly with an increasing virtual ESS. This finding gives a first hint that a prior with a large ESS might be beneficial for the MSP learning principle, while we cannot observe a similar effect for the MAP learning principle in this case.

Next, we consider the lines $\beta_1 = \nu - \beta_0$, which correspond to the hybrid learning principles GDT and PGDT for $\nu = 1$ and $\nu = 0.5$, respectively. For the GDT learning principle, the sensitivity ranges from 54.7% for $\underline{\beta}$ = (0, 1, 0) to 55.2% for $\underline{\beta}$ = (1, 0, 0), reaching a maximum of 56.9% for $\underline{\beta}$ = (0.55, 0.45, 0). For the PGDT learning principle, the sensitivity ranges from 54.9% for $\underline{\beta}$ = (0, 0.5, 0.5) to 55.6% for $\underline{\beta}$ = (0.5, 0, 0.5), reaching a maximum of 57.1% for $\underline{\beta}$ = (0.3, 0.2, 0.5). For both learning principles, we find that the sensitivity is initially increasing and finally decreasing. This observation indicates that neither the MAP nor the MSP learning principle with a Generalized Dirichlet prior representing uniform pseudo data is optimal for estimating the parameter vector $\underline{\lambda}$.

Next, we investigate the interior of the simplex. We vary both β_0 and β_1 along a grid with step width 0.05, and we find the highest sensitivity of 57.3% for $\underline{\beta}$ = (0.1, 0.1, 0.8). We find the region of highest sensitivity clearly inside the simplex near the angle bisector. This region corresponds to the MSP learning principle with an informative prior based on weighted likelihood and weighted original prior. Comparing the highest sensitivity for the GDT, the PGDT, and the unified generative-discriminative learning principle, we find that it increases from 56.9% over 57.1% to 57.3%, confirming that the prior can have a positive influence on the performance.

Turning to the results of the other three TFs GATA, NF-κB, and Thyroid, we find qualitatively similar results. The highest sensitivities are located inside the simplex, while the lowest sensitivities are located on the axes. For BSs of the TF GATA, we obtain a sensitivity of 77.5% for $\underline{\beta}$ = (0.45, 0.25, 0.3), for the BSs of the TF NF-κB, we obtain a sensitivity of 81.8% for $\underline{\beta}$ = (0.4, 0.55, 0.05), and for the BSs of the TF Thyroid, we obtain 52.3% for $\underline{\beta}$ = (0.4, 0.55, 0.05). Similar to the data set of AR/GR/PR, we find a small region with high sensitivity for the BSs of the TFs NF-κB and Thyroid, while we find a broad region with high sensitivity for the BSs of the TF GATA.

We summarize the sensitivities for the ML, the MCL, the MAP, the MSP, and the unified generative-discriminative learning principle in Table 2. We find that for all four TFs the unified generative-discriminative learning principle yields the highest sensitivities. Regarding the β_1-axis, which corresponds to the MAP learning principle using the Generalized Dirichlet prior representing uniform pseudo data with different ESSs, we find that increasing the prior weight β_2, which is equivalent to decreasing the generative weight β_1, often reduces the sensitivity. We obtain the lowest sensitivity for the MAP learning principle for the largest prior weights β_2 in almost all cases. In contrast to this observation, we find on the β_0-axis, which correspond to the MSP learning principle with the Generalized Dirichlet prior representing uniform pseudo data with different ESSs, that increasing the prior weight β_2 improves the sensitivity at least initially.

Interestingly, we obtain qualitatively similar results when using other performance measures (Appendix B in Additional File 1). These observations suggest that the same classifier trained either by generative or by discriminative learning principles may prefer different ESSs even if one uses a prior that corresponds to uniform pseudo data. Hence, the strength of the prior has a decisive influence on comparisons of the results from generative and discriminative learning principles as well as the results of Bayesian hybrid learning principles as for instance PGDT learning principle. Most importantly, we find that the unified generative-discriminative learning principle leads to an improvement for almost all of the studied data sets and performance measures.

Table 2 Results for four data sets

	AR/GR/PR	GATA	NF-κB	Thyroid
ML	54.7	77.0	81.6	51.3
MCL	55.2	73.2	76.5	50.0
MAP	55.1	77.0	81.6	51.3
MSP	56.9	77.0	79.6	50.4
Unified	**57.3**	**77.5**	**81.8**	**52.3**

Summary the results of Figure 2 for the 4 data sets containing the highest sensitivity for the ML, the MCL, the MAP, the MSP, and the unified generative-discriminative learning principle. For the MAP, the MSP, and the unified generative-discriminative learning principle, we present the best results form the simplex $\underline{\beta}$ which correspond to one of these learning principles (see Figure 1b). For each data set, the highest sensitivity is displayed in bold.

Conclusions

A plethora of algorithms for the recognition of short DNA sequence motifs has been proposed in the last decades. These algorithms differ by their underlying statistical models and the employed learning principles. In bioinformatics, generative learning principles have a long tradition, but recently it was shown that discriminative learning principles can lead to an improvement of the recognition of short signal sequences.

We introduce a unified generative-discriminative learning principle that contains the ML, the MAP, the MCL, the MSP, the GDT, and the PGDT learning principle as limiting cases. This learning principle interpolates between the likelihood, the conditional likelihood, and the prior, spanning a three-dimensional simplex, which allows a more detailed comparison of different learning principles. Furthermore, we find that under mild assumptions each point on the simplex can be interpreted as MSP learning principle using an informative prior composed of a weighted likelihood and a weighted original prior.

We find that the unified generative-discriminative learning principle improves the performance of classifiers for the recognition of vertebrate TFBSs over any of the six established learning principles it contains as special case. We make all implementations available for the scientific community as part of the open-source Java library Jstacs [45], which allows using this learning principle easily for other bioinformatics problems. Although we demonstrate the utility of the unified generative-discriminative learning principle only for four data sets of TFBSs and four performance measures, it is conceivable that it can be successfully applied to other multinomial data such as data of transcription start sites, donor and acceptor splice sites, splicing enhancers and silencers, as well as binding sites of insulators, nucleosomes, and miRNAs, as well as continuous data.

Methods

Considering the task of determining the optimal parameter vector $\hat{\underline{\lambda}}$, we find that generative learning principles often allow to estimate $\hat{\underline{\lambda}}$ analytically for simple models such as Markov models, but one must use numerical optimization procedures for discriminative and hybrid learning principles, and consequently for the unified generative-discriminative learning principle as well. If the conditional likelihood, the likelihood, and the prior are log-convex functions, we can use any numerical algorithm to determine the globally optimal parameter vector $\hat{\underline{\lambda}}$ for the unified generative-discriminative learning principle.

Different numerical methods including steepest descent, conjugate gradient, quasi-Newton methods, and limited-memory quasi-Newton methods have been evaluated in [46]. In the case studies presented in the previous subsection, we use a limited-memory quasi-Newton method. In analogy to [37], we fix $\underline{\beta}$ for the unified generative-discriminative learning principle, and we compute the results for a grid of given values of $\underline{\beta}$, providing an overall impression of the performance for the whole simplex $\underline{\beta}$.

The unified generative-learning principle can in principle be used for all types of data, and it is not limited to multinomial data presented in section *Testing*. We make all implementations available for the scientific community as part of the open-source Java library Jstacs [45]. Jstacs comprises an efficient representation of sequence data and provides object-oriented implementations of many statistical models. We implement the unified generative-discriminative learning principle as a multi-threaded class based on the Jstacs class hierarchy [47]. This allows applying the learning principle efficiently on multi-core computers and to other statistical models. For optimizing parameters, we use optimization procedures provided by Jstacs.

Availability and Requirements

Project name: GenDisMix
Project home page: [47]
Operating system(s): Platform independent
Programming language: Java 1.5
Requirements: Jstacs 1.3
License: GNU General Public License version 3

Additional file 1: Appendix. This file contains additional information about Markov random fields and the case studies.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-98-S1.PDF]

List of abbreviations used

BS: binding site; ESS: equivalent sample size; GDT: generative-discriminative trade-off; MAP: maximum a posteriori; MCL: maximum conditional likelihood; ML: maximum likelihood; MSP: maximum supervised posterior; PGDT: penalized generative-discriminative trade-off; PWM: position weight matrix; TF: transcription factor; TFBS: transcription factor binding sites; WAM: weight array matrix.

Acknowledgements

We thank Alexander Zien for helpful discussions and four anonymous reviewers for their valuable comments. This work was supported by grant XP3624HP/0606T by the Ministry of Culture of Saxony-Anhalt.

Author details

[1]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany. [2]Institute of Computer Science, Martin Luther University Halle-Wittenberg, Germany.

Authors' contributions

JK and IG developed the basic ideas. JK and JG implemented the software. JK performed the case studies. All authors contributed to data analysis, writing, and approved the final manuscript.

Received: 5 November 2009 Accepted: 22 February 2010
Published: 22 February 2010

References

1. Kel AE, Gössling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E: **MATCH: A tool for searching transcription factor binding sites in DNA sequences.** *Nucleic Acids Res* 2003, **31(13)**:3576-3579.
2. Barash Y, Elidan G, Friedman N, Kaplan T: **Modeling Dependencies in Protein-DNA Binding Sites.** *In proceedings of Seventh Annual International Conference on Computational Molecular Biology* 2003, 28-37.
3. Sonnenburg S, Zien A, Rätsch G: **ARTS: accurate recognition of transcription starts in human.** *Bioinformatics* 2006, **22(14)**:e472-e480.
4. Abeel T, Peer Van de Y, Saeys Y: **Toward a gold standard for promoter prediction evaluation.** *Bioinformatics* 2009, **25(12)**:i313-i320.
5. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA.** *J Mol Biol* 1997, **268**:78-94.
6. Salzberg SL: **A method for identifying splice sites and translational start sites in eukaryotic mRNA.** *Comput Appl Biosci* 1997, **13(4)**:365-376.
7. Yeo G, Burge CB: **Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals.** *Journal of Computational Biology* 2004, **11(2-3)**:377-394.
8. Segal E, Fondufe-Mittendorf Y, Chen L, Thåström A, Field Y, Moore IK, Wang JPZ, Widom J: **A genomic code for nucleosome positioning.** *Nature* 2006, **442(7104)**:772-778.
9. Peckham HE, Thurman RE, Fu Y, Stamatoyannopoulos JA, Noble WS, Struhl K, Weng Z: **Nucleosome positioning signals in genomic DNA.** *Genome Res* 2007, **17(8)**:1170-1177.
10. Lewis BP, hung Shih I, Jones-Rhoades MW, Bartel DP, Burge CB: **Prediction of Mammalian MicroRNA Targets.** *Cell* 2003, **115(7)**:787-798.
11. Maragkakis M, Reczko M, Simossis VA, Alexiou P, Papadopoulos GL, Dalamagas T, Giannopoulos G, Goumas G, Koukis E, Kourtis K, Vergoulis T, Koziris N, Sellis T, Tsanakas P, Hatzigeorgiou AG: **DIANA-microT web server: elucidating microRNA functions through target prediction.** *Nucl Acids Res* 2009, **37(suppl 2)**:W273-276.
12. Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, Green RD, Zhang MQ, Lobanenkov VV, Ren B: **Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome.** *Cell* 2007, **128(6)**:1231-1245.
13. Staden R: **Computer methods to locate signals in nucleic acid sequences.** *NAR* 1984, **12**:505-519.
14. Stormo G, Schneider T, Gold L, Ehrenfeucht A: **Use of the 'perceptron' algorithm to distinguish translational initiation sites.** *NAR* 1982, **10**:2997-3010.
15. Zhang M, Marr T: **A weight array method for splicing signal analysis.** *Comput Appl Biosci* 1993, **9(5)**:499-509.
16. Yakhnenko O, Silvescu A, Honavar V: **Discriminatively Trained Markov Model for Sequence Classification.** *ICDM '05: Proceedings of the Fifth IEEE International Conference on Data Mining, Washington, DC, USA: IEEE Computer Society* 2005, 498-505.
17. Keilwagen J, Grau J, Posch S, Grosse I: **Recognition of splice sites using maximum conditional likelihood.** *LWA: Lernen - Wissen - Abstraktion* Hinneburg A 2007, 67-72.
18. Cai D, Delcher A, Kao B, Kasif S: **Modeling splice sites with Bayes networks.** *Bioinformatics* 2000, **16(2)**:152-158.
19. Ben-Gal I, Shani A, Gohr A, Grau J, Arviv S, Shmilovici A, Posch S, Grosse I: **Identification of transcription factor binding sites with variable-order Bayesian networks.** *Bioinformatics* 2005, **21(11)**:2657-2666.
20. Culotta A, Kulp D, McCallum A: **Gene Prediction with Conditional Random Fields.** *Tech Rep Technical Report UM-CS-2005-028* University of Massachusetts, Amherst 2005.
21. Bernal A, Crammer K, Hatzigeorgiou A, Pereira F: **Global discriminative learning for higher-accuracy computational gene prediction.** *PLoS Comput Biol* 2007, **3(3)**:e54.
22. Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavesi G, Pesole G, Régnier M, Simonis N, Sinha S, Thijs G, van Helden J, Vandenbogaert M, Weng Z, Workman C, Ye C, Zhu Z: **Assessing computational tools for the discovery of transcription factor binding sites.** *Nature Biotechnology* 2005, **23**:137-144.
23. Ng AY, Jordan MI: **On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes.** *Advances in Neural Information Processing Systems* Cambridge, MA: MIT PressDietterich T, Becker S, Ghahramani Z 2002, **14**:605-610 [http://citeseer.ist.psu.edu/542917.html].
24. Greiner R, Su X, Shen B, Zhou W: **Structural Extension to Logistic Regression: Discriminative Parameter Learning of Belief Net Classifiers.** *Machine Learning Journal* 2005, **59(3)**:297-322.
25. Pernkopf F, Bilmes JA: **Discriminative versus generative parameter and structure learning of Bayesian network classifiers.** *Proceedings of the 22nd International Conference on Machine Learning* 2005, 657-664.
26. Grau J, Keilwagen J, Kel A, Grosse I, Posch S: **Supervised posteriors for DNA-motif classification.** *German Conference on Bioinformatics, Lecture Notes in Informatics (LNI) - Proceedings* Gesellschaft für Informatik (GI)Falter C, Schliep A, Selbig J, Vingron M, Walter D 2007, 123-134.
27. Keilwagen J, Grau J, Posch S, Grosse I: **Apples and oranges: avoiding different priors in Bayesian DNA sequence analysis.** *BMC Bioinformatics* 2009.
28. Fisher RA: **On the Mathematical Foundations of Theoretical Statistics.** 1922.
29. Aldrich J: **R. A. Fisher and the Making of Maximum Likelihood 1912-1922.** *Statistical Science* 1997, **12(3)**:162-176.
30. Bishop CM: *Pattern Recognition and Machine Learning* Springer 2006.
31. Redhead E, Bailey TL: **Discriminative motif discovery in DNA and protein sequences using the DEME algorithm.** *BMC Bioinformatics* 2007, **8**:385.
32. Wettig H, Grünwald P, Roos T, Myllymäki P, Tirri H: **On Supervised Learning of Bayesian Network Parameters.** *Tech Rep HIIT Technical Report 2002-1* Helsinki Institute for Information Technology HIIT 2002 [http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.3.9589].
33. Grossman D, Domingos P: **Learning Bayesian network classifiers by maximizing conditional likelihood.** *ICML* ACM Press 2004, 361-368.
34. Feelders A, Ivanovs J: **Discriminative Scoring of Bayesian Network Classifiers: a Comparative Study.** *Proceedings of the third European workshop on probabilistic graphical models* 2006, 75-82.
35. Grüwald P, Kontkanen P, Myllymäki P, Roos T, Tirri H, Wettig H: **Supervised posterior distributions.** *Presented at the Seventh Valencia International Meeting on Bayesian Statistics* 2002.
36. Cerquides J, de Mántaras RL: **Robust Bayesian Linear Classifier Ensembles.** *ECML* 2005, 72-83.
37. Bouchard G, Triggs B: **The Tradeoff Between Generative and Discriminative Classifiers.** *IASC International Symposium on Computational Statistics (COMPSTAT), Prague* 2004, 721-728 [http://lear.inrialpes.fr/pubs/2004/BT04].
38. Raina R, Shen Y, Ng AY, McCallum A: **Classification with Hybrid Generative/Discriminative Models.** *Advances in Neural Information Processing Systems 16* Cambridge, MA: MIT PressThrun S, Saul L, Schölkopf B 2004.
39. Lasserre JA, Bishop CM, Minka TP: **Principled Hybrids of Generative and Discriminative Models.** *Proc IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2006, **1**:87-94 [http://research.microsoft.com/en-us/um/people/cmbishop/downloads/bishop-cvpr-06.pdf].
40. Mccallum A, Pal C, Druck G, Wang X: **Multi-conditional learning: Generative/discriminative training for clustering and classification.** *NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE* 2006, 433-439 [http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.67.5681].
41. Bouchard G: **Bias-Variance Tradeoff in Hybrid Generative-Discriminative Models.** *ICMLA '07: Proceedings of the Sixth International Conference on Machine Learning and Applications* Washington, DC, USA: IEEE Computer Society 2007, 124-129.
42. Xue JH, Titterington DM: **Interpretation of hybrid generative/discriminative algorithms.** *Neurocomputing* 2009, **72(7-9)**:1648-1655.
43. Hastie T, Tibshirani R, Friedman JH: *The elements of statistical learning: data mining, inference, and prediction* Springer 2009 [http://www-stat.stanford.edu/~hastie/Papers/ESLII.pdf].
44. Wingender E, Dietze P, Karas H, Knuppel R: **TRANSFAC: A database on transcription factors and their DNA binding sites.** *Nucleic Acids Res* 1996, **24**:238-241.
45. **A Java framework for statistical analysis and classification of biological sequences.** [http://www.jstacs.de/].
46. Wallach H: **Efficient Training of Conditional Random Fields.** *Master's thesis* University of Edinburgh 2002.
47. **Jstacs Projects: GenDisMix.** [http://www.jstacs.de/index.php/GenDisMix].

doi:10.1186/1471-2105-11-98
Cite this article as: Keilwagen *et al.*: **Unifying generative and discriminative learning principles.** *BMC Bioinformatics* 2010 **11**:98.