BIOINFORMATICS

Vol. 00 no. 00 xxxx
Pages 1–8

LOGICOIL — Multi-state prediction of coiled-coil oligomeric state

Thomas. L. Vincent [1,2], Peter J. Green [3] and Derek N. Woolfson [1,4] *

[1]School of Chemistry, University of Bristol, Bristol, BS8 1TS.
[2]Bristol Centre for Complexity Science, University of Bristol, Bristol, BS8 1TR.
[3]School of Mathematics, University of Bristol, Bristol, BS8 1TW.
[4]School of Biochemistry, Medical Sciences Building, University of Bristol, Bristol, BS8 1TD.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: The coiled coil is a ubiquitous α-helical protein-structure domain that directs and facilitates protein-protein interactions in a wide variety of biological processes. At the protein-sequence level, the coiled coil is readily recognized via a conspicuous heptad repeat of hydrophobic and polar residues. However, structurally coiled coils are more complicated, existing in a wide range of oligomer states and topologies. As a consequence, predicting these various states from sequence remains an unmet challenge.

Results: This work introduces LOGICOIL, the first algorithm to address the problem of predicting multiple coiled-coil oligomeric states from protein-sequence information alone. By covering $> 90\%$ of the known coiled-coil structures, LOGICOIL is a net improvement compared to other existing methods, which achieve a predictive coverage of $\sim 31\%$ of this population. This leap in predictive power offers better opportunities for genome-scale analysis, and analyses of coiled-coil containing protein assemblies.

Availability: LOGICOIL is available via a web-interface at http://coiledcoils.chm.bris.ac.uk/LOGICOIL. Source code, training sets and Supporting Information can be downloaded from the same site.

Contact: D.N.Woolfson@bristol.ac.uk

1 INTRODUCTION

Coiled coils are protein-structure domains that comprise two or more alpha-helices that wrap around each other, typically in a left-handed fashion, and which interact through specific packing interactions known as knobs-into-hole packing. (Crick, 1953; Lupas and Gruber, 2005). While accounting for approximately 2.9% (range, 0.3% – 6.5%) of the protein-encoding regions of genes (Rackham *et al.*, 2010), coiled coils are also actively involved in the mediation of protein-protein interactions across a wide array of biological functions; from transcription, through membrane remodeling, to cell and tissue structure and stability (Yu, 2002). Despite its functional diversity, the coiled coil is characterized by a straightforward sequence motif of hydrophobic (H) and polar (P) residues. The positions within this HPPHPPP motif, referred to as the heptad repeat, are typically labelled a through g, with hydrophobic residues generally occupying the a and d positions, and polar residues falling at the other positons. Given this common sequence pattern, the 3-dimensional structures adopted by naturally occurring coiled coils display a remarkable diversity. Applying SOCKET, an algorithm that finds knobs-into-holes packing interactions within structurally resolved proteins (Walshaw and Woolfson, 2001b), to the RCSB PDB (Berman *et al.*, 2000) reveals an abundance of coiled-coil architectures and topologies (Testa *et al.*, 2009). Indeed, coiled-coil assemblies have been shown to contain different numbers of helices of parallel or anti-parallel orientation, that may be formed from the same (homo) or different (hetero) helical sequences (Moutevelis and Woolfson, 2009; Lupas and Gruber, 2005).

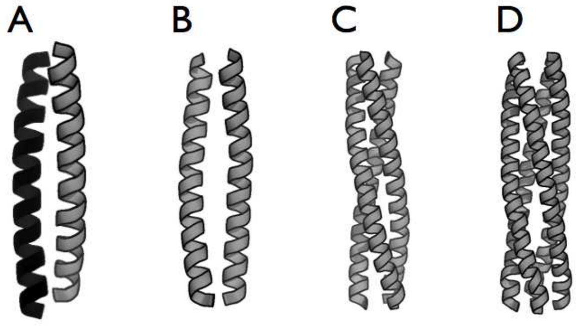

Fig. 1. Different oligomeric states that may be adopted by a coiled-coil structure and that are targeted by LOGICOIL: antiparallel dimer (A); parallel dimer (B); trimer (C) and tetramer (D). While other coiled-coil topologies are observed in nature, the four oligomeric states displayed here account for over 90% of the known population.

Here, we focus on the prediction of coiled-coil oligomeric state. Six algorithms exist to tackle this problem: SCORER (Woolfson and Alber, 1995), which has been recently redefined and retrained in SCORER 2.0 (Armstrong *et al.*, 2011), and PrOCoil (Mahrenholz *et al.*, 2011) achieve high success rates when separating coiled-coil sequences, but these methods are strictly limited to the

*To whom correspondence should be addressed.

discrimination of parallel dimeric and parallel trimeric coiled-coil structures. Multicoil2 (Trigg *et al.*, 2011) and its predecessor MultiCoil (Wolf *et al.*, 1997) follow a different approach to predict both the location and oligomeric state of coiled coils in protein sequences. However, their oligomeric state functions remains limited to the discrimination of parallel dimers and trimers.

Thus, these algorithms cover only a small subset of the known coiled-coil structural space, limiting their usefulness. For example, antiparallel dimers, so far excluded from all prediction analysis, account for well over 50% of the total coiled-coil structure population and represent a wealth of untapped data (Moutevelis and Woolfson, 2009). Current *de novo* methods cover only $\sim 31\%$ of the total coiled-coil population. Thus, simple inclusion of antiparallel dimeric and tetrameric structures would increase coverage to over 90%.

Homology based approaches such as SPIRICOIL (Rackham *et al.*, 2010) partially accomplish the task of multi-state coiled-coil oligomeric state prediction, but cannot be used to classify the oligomeric state of coiled-coil sequences *ab initio*; *i.e.*, those without structurally defined precedents. As a consequence, we regard the development of a *ab initio* multi-state predictor to be the next logical step in coiled-coil structure analysis and prediction. To the best of our knowledge, no work has yet treated the *ab initio* problem of multi-state classification of coiled-coil oligomers. As previous attempts have focussed on two-state predictions and therefore been been tailored towards binary response problems, they could not be systematically extended for the purpose of multi-state classification. As such, the problem required the development of statistical techniques and algorithm capable of discriminating between multiple coiled-coil oligomeric states.

2 APPROACH

LOGICOIL is based on the simultaneous use of Bayesian variable selection, and multinomial probit regression for prediction. We favour this methodology over others for its ability to perform variable selection and parameter estimation simultaneously. Furthermore, the Bayesian paradigm allowed us to obtain informative posterior distributions on the selected parameters, while providing a convenient framework for the use of prior information based on biological data and expert knowledge. Although other commonly used methods for classification such as support vector machines can incorporate variable selection in the context binary classification problems (Becker *et al.*, 2009; Hochreiter and Obermayer, 2006), these methods are not yet applicable to multi-class problems. The statistical framework used for LOGICOIL can be easily applied to the prediction of multiple coiled-coil oligomeric states, while accounting for the inclusion of higher-order associations, such as intrahelical pairwise residue associations. Pairwise interaction effects have been included in other algorithms that aim to predict coiled-coil oligomeric state, but these are limited to two-state predictors and have yet be extended to multinomial classification (Wolf *et al.*, 1997; Mahrenholz *et al.*, 2011). The higher-dimensional models and associated computational challenges that this approach usually may entail are becoming increasingly accessible; notably through the use of Markov Chain Monte Carlo (MCMC) methods and increased computational power. After discussing multinomial regression models in the Bayesian context, we describe a Bayesian variable selection scheme that selects the most relevant pairwise associations. Furthermore, we discuss computational and implementation issues of the variable selection scheme and how they were handled.

3 BAYESIAN MULTINOMIAL PROBIT REGRESSION WITH VARIABLE SELECTION

3.1 Problem Formulation

Assume a data set $\{y_i; x_{i1}, \ldots, x_{ip}\}_{i=1}^{n}$, where n is the number of observed samples, $y_i \in \{1, \ldots, C\}$ is a polytomous outcome and x_{ij} is the observed value of the j^{th} predictor in the i^{th} sample with $j = 1, 2, \ldots, p$. We also denote the predictor matrix $\mathbf{X} = (x_{ij})_{n,p}$ as:

$$\mathbf{X} = \begin{bmatrix} \text{predictor 1} & \text{predictor 2} & \cdots & \text{predictor } p \\ x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \qquad (1)$$

The relationship between the probability $\Pr(y_i = c|X_i) = \pi_{ic}$ of observing response c given X_i can be conveniently modeled using Generalized Linear Models (McCullagh and Nelder, 1989), and is determined by a $(1\times p)$ matrix of regression coefficients $\boldsymbol{\beta}_c$. Although the unknown parameters in $\boldsymbol{\beta}_c$ can be estimated through maximum likelihood techniques, data augmentation techniques and Gibbs sampling can also be used to allow for efficient simulation under the Bayesian paradigm (Albert and Chib, 1993; Holmes and Held, 2006). Here, we consider a Bayesian auxiliary variable model that assumes the probabilities π_{ic} to be related to X_i and $\boldsymbol{\beta}_c$ through a probit link function.

The key idea proposed by (Albert and Chib, 1993) is to generate the y_i from the X_i through latent vectors $\mathbf{Z}_i = (z_{i,1}, \ldots, z_{i,C-1})$, which are thresholded to determine the y_i; this means that probabilities π_{ic} are defined only implicitly. The approach is expressed precisely through the following notation:

$$\begin{aligned} y_i &= \begin{cases} c, & \text{if } z_{i,c} = max\{\mathbf{Z}_i, 0\} \\ C, & \text{if } max\{\mathbf{Z}_i\} \leq 0, \end{cases} \\ z_{i,c} &= \eta_{i,c} + \epsilon_{i,c} \\ \eta_{i,c} &= x_i\beta_c \\ \epsilon_{i,c} &\sim N(0,1) \\ \beta_c &\sim \pi(\beta_c) \end{aligned} \qquad (2)$$

where y_i is now conditional on the newly introduced auxiliary variable vector $\mathbf{Z}_i$.

Writing $\mathbf{z}_c = [z_{1,c}, \ldots, z_{n,c}]^T$ and $\boldsymbol{\epsilon}_c = [\epsilon_{1,c}, \ldots, \epsilon_{n,c}]^T$, the auxiliary variable can be expressed in vector form as:

$$\mathbf{z}_c = \mathbf{X}\boldsymbol{\beta}_c + \boldsymbol{\epsilon}_c, \qquad \boldsymbol{\epsilon}_c \sim N(0, \Sigma) \text{ for } c = 1, \ldots, C-1. \qquad (3)$$

where $\mathbf{X}$ is the $(C-1) \times p$ predictor matrix, $\boldsymbol{\beta}_c$ is the $1 \times p$ matrix of fixed coefficients with respect to the response c, ϵ_i is a $(C-1) \times 1$ vector of errors and Σ is a $(C-1) \times (C-1)$ positive definite matrix with $\sigma_{11} = 1$.

3.2 Bayesian variable selection

The Bayesian multinomial probit model in (2) is well-suited to variable selection problems. Throughout the past decade, a number of special forms of the reversible jump sampler introduced by (Green, 1995) have been advanced. In particular, much work has focussed on data augmented models that propose efficient predictor selection methods (Stingo and Vanucci, 2010; Sha *et al.*, 2004; Zhou *et al.*, 2004, 2006; Tuchler, 2008; Holmes and Held, 2006; Ai-Jun and Xin-Yuan, 2010; Gustafson and Lefebvre, 2008). Here, we use the scheme described in (Holmes and Held, 2006) and (Gustafson and Lefebvre, 2008). Their approach introduces a covariate indicator vector that determines if predictors are in or out of the model:

$$\gamma_i = \begin{cases} 1, \text{ if } \beta_j \neq 0 \text{ (the j-th predictor is selected)}, \\ 0, \text{ if } \beta_j = 0 \text{ (the j-th predictor is not selected)}. \end{cases} \quad (4)$$

The parameter γ is then included in (3) so that:

$$\mathbf{z}_c = \mathbf{X}_\gamma \boldsymbol{\beta}_{c,\gamma} + \mathbf{e}_c, \qquad c = 1, \ldots, C-1. \quad (5)$$

where $\mathbf{X}_\gamma$ are the elements of $\mathbf{X}$ set to 1 and $\boldsymbol{\beta}_{c,\gamma}$ consists of all the non-zero elements of $\boldsymbol{\beta}_c$. For the Bayesian variable selection scheme in (5), the remaining task involved the estimation of the indicator vector $\boldsymbol{\gamma} = \{\gamma_1, \ldots, \gamma_p\}$ and the corresponding $\boldsymbol{\beta}_{\gamma,c}$ and $\mathbf{z}_c$.

A Gibbs sampler is employed to estimate all the parameters in the model. Following from others, we choose a "ridge" prior $\pi(\boldsymbol{\beta}) = N_p(0, v\mathbf{I}_p)$ on the $(C-1) \times p$ matrix of parameters $\boldsymbol{\beta}$, where $N_p(\mu, \hat{\boldsymbol{\Sigma}})$ represents a p-multivariate normal distribution with mean μ and covariance matrix $\hat{\boldsymbol{\Sigma}}$, and $\mathbf{I}_p$ is the $p \times p$ identity matrix (Sha *et al.*, 2004; Brown *et al.*, 2002). It should be noted that alternative choices of prior distribution also exist (Liang *et al.*, 2008; O'Hara and Sillanpaaa, 2009). The detailed derivation of the posterior distributions of the parameters and the Gibbs sampler used to estimate the parameter vectors $\boldsymbol{\gamma}$, $\boldsymbol{\beta}_{\gamma,c}$ and $\mathbf{z}_c$ for Bayesian variable selection are identical to that of Holmes & Held and Gustafson & Lefebvre (Holmes and Held, 2006; Gustafson and Lefebvre, 2008).

3.3 Computing the Response Probabilities

For the development of LOGICOIL, we treated pairwise-association selection and coiled-coil-oligomeric-state classification in two sequential steps. Pairwise associations were first selected using Bayesian variable selection. From the $\{\gamma^{(t)}, \beta_c^{(t)}, z_c^{(t)}, t = 1, \ldots, T\}$ MCMC samples obtained from the Gibbs sampling scheme, the pairwise associations with the highest posterior probability of inclusion were assumed to play the strongest role in predicting the target coiled-coil oligomeric state. Posterior inclusion probabilities were computed according to:

$$p(\gamma_i = 1) = \frac{1}{T}\sum_{t=1}^{T} \gamma_i^{(t)} \quad (6)$$

where

$$\gamma_i^{(t)} = \begin{cases} 1, \text{ if } \gamma_i \text{ was included in the model at the } t^{th} \text{ iteration} \\ 0, \text{ otherwise} \end{cases}$$

Once the strongest pairwise associations had been identified, the parameter values of the retained predictors were fitted in a separate model. Here, multinomial probit regression was chosen to preserve consistency with the variable-selection process. There is no closed form for the likelihood function of multinomial probit models but practical MCMC methods have been proposed to compute parameter estimates. Here, parameter values in the multinomial probit model were estimated using the *MNL* library available in the R software (Imai and van Dyk, 2005b,a; Team, 1993).

From the T samples $\{\beta_c^{(t)}, z_c^{(t)}, t = 1, \ldots, T\}$ of the secondary MCMC scheme, the probability of a given test coiled-coil sequence under each class was computed as:

$$p(y_i = c|X_i) = \frac{1}{T}\sum_{t=1}^{T}\Big(\Phi(X_i\beta_c^{(t)}) = \max\{\Phi(X_i\beta_j^{(t)}), 0)\}\Big)$$
$$j = 1, \ldots, C-1 \text{ and } j \neq c$$
$$p(y_i = C|X_i) = \frac{1}{T}\sum_{t=1}^{T}\Big(\max\{\Phi(X_i\beta_j^{(t)})\} \leqslant 0\Big)$$
$$j = 1, \ldots, C-1 \quad (7)$$

where $\Phi(.)$ is the multivariate normal distribution and X_i is the $(C-1) \times p$ dimensional vector for observation i.

4 PRACTICAL IMPLEMENTATION

4.1 Convergence of MCMC sampling schemes

For all the examples discussed below, the convergence of the MCMC sampling schemes was assessed using multiple chains with different random number seeds and starting values. The number of iterations necessary for the MCMC sampling scheme to converge, otherwise known as the burn-in, was estimated at the point for which the independent MCMC runs displayed similar values. Autocorrelation in the MCMC samples was also checked to ensure that the chains were mixing adequately. Once a suitable burn-in period had been identified, a single, long chain was ran and used to compute parameter estimates and posterior probabilities. In line with accepted guidelines, it was also ensured that the acceptance ratio of the MH step was kept in the range of [25% – 45%] during simulations (Gelman *et al.*, 2004).

4.2 Preselection of variables

Bayesian variable selection provided a rigorous framework to select the strongest predictor variables in a model, but was computationally intensive given the large number of variables in the model. To facilitate simulations and reduce the dimensions of our data, a preselection filter was used to choose a smaller subset of variables, to which Bayesian variable selection was subsequently applied.

Let $n(a_1 \rightarrow r_1, a_2 \rightarrow r_2, c)$ be the number of times the following is observed: amino acid a_1 at register position r_1, and amino acid a_2 at register position r_2, and response c in the dataset; antiparallel dimer, parallel dimer, trimer or tetramer. Using the hypergeometric distribution, the probability of observing the spatial interaction $n(a_1 \rightarrow r_1, a_2 \rightarrow r_2, c)$ exactly k times can be written as:

$$p(n(a_1 \rightarrow r_1, a_2 \rightarrow r_2, c) = k) = \frac{n(a_1 \rightarrow r_1, + \rightarrow r_2, c)n(+ \rightarrow r_1, a_2 \rightarrow r_2, c)}{n(+,+,c)} \quad (8)$$

where the symbol $+$ denotes summing out over all other amino acids and $c \in \{1, \ldots, C\}$. A two-sided p-value was assigned to each observation and used to preselect the most significant pairwise associations. To avoid missing important predictor variables, a relaxed preselection criterion was taken by setting a p-value threshold of 0.1.

4.3 Prior Distribution

The posterior inclusion probabilities $p(\gamma_i = 1|\mathbf{Y}, \mathbf{X})$ have been reported to be sensitive to the hyperparameter v in the prior distribution $\pi(\boldsymbol{\beta}) = N_p(0, v\mathbf{I}_p)$ (Lamnisos *et al.*, 2010, 2009; Fernandez *et al.*, 2001). To ensure we selected the variables that helped achieve optimal out-of-sample predictive performance, we followed the methodology adopted in (Lamnisos *et al.*, 2010) and used K-fold cross-validation across a range of values on the hyperparameter prior v. The K-fold predictive scores were estimated at $l = 20$ values of v equally spaced in the logarithmic scale with lower value 0.1 and upper value 500. Since this cross-validation methodology requires Kl MCMC runs to estimate the K partition scores for the l values of v, we use a value of $K = 10$ to contain computational expenses. Similar predictive performance were obtained for values of v in the interval $(0.1, 59)$, which is in accordance with the guideline range of v proposed in Sha *et al.* (2004). For the chosen values of v, residue pairs were preselected according to the method described in §3.5 and subsequently applied to Bayesian variable selection model.

5 METHODS

5.1 Coiled-coil training and test sets

The sequences of antiparallel dimeric, and parallel dimeric, trimeric and tetrameric canonical — that is, heptad based — coiled coils longer than 14 residues in length were obtained from the CC+ database (Testa *et al.*, 2009). All coiled-coil sequences were aligned using Clustalw2 (Larkin *et al.*, 2007) (maximum gap penalties were used to conserve the alignment of the heptad repeat), and then culled using CD-HIT (Li and Godzik, 2006) at a redundancy cutoff of 50%. The corresponding structures were validated by eye to ensure that all remaining sequences belonged to well-defined coiled-coil systems. We use a 50% identity threshold rather than $25-30\%$, often used for culling protein datasets, as we find the latter to be too restrictive for coiled-coil sequences, which have a constricted amino acid usage, and therefore regarded as regions of low complexity (Armstrong *et al.*, 2011; Jones and Swindells, 2002). The final dataset — referred to as the pristine dataset — comprises 670 antiparallel dimeric, 173 parallel dimeric, 55 trimeric and 39 tetrameric coiled-coil sequences.

5.2 Assessing predictive performance

For problems involving binary response classes, a popular method to quantify the prediction accuracy of a classifier during cross-validation is the Receiver Operator Characteristic (ROC) curve, and the associated Area Under the Curve (AUC) measure (Fawcett, 2006). The overall performance, denoted as mAUC, of a multi-class classifier can be computed using a generalization of the AUC for multiple class classification problems as defined in (Hand and Till, 2001). By averaging the AUC obtained for each pair of classes, the mAUC discrimination rate can be obtained through the score:

$$\text{mAUC} = \frac{2}{c(c-1)} \sum_{i<j} AUC_{c_i,c_j} \qquad (9)$$

where $\text{AUC}_{c_i,c_j}, (c_i, c_j) \in 1, \ldots, C$ with $(c_i \neq c_j)$ is the probability that a randomly drawn member from response class i will have a lower estimated probability of belonging to class j than a randomly drawn member of class j. Although no equivalent to ROC curves exists to plot multi-class performance measures, the pairwise AUC's can be visualized in the form on a spider-web diagram. This has the advantage of showing the achieved discrimination rate between all pairs of classes, therefore allowing for the identification of pairs of classes that are well separated and those that are not, regardless of the reported overall mAUC score. All AUC values were computed using the ROCR and caTools packages in the R software (Sing *et al.*, 2005).

6 MULTI-STATE PREDICTION OF COILED-COIL OLIGOMERIC STATE.

6.1 Including pairwise associations in the model

Prior to running the Bayesian variable selection scheme on a larger scale, we performed a trial run on a limited set of pairwise associations. Due to its well-documented and important role in oligomer formation, we implemented Bayesian variable selection to detect the strongest residue $d - a$ pairs, which is equivalent to a $i \rightarrow i + 4$ association (Figure 2). The *d* and *a* register positions are part of the coiled-coil hydrophobic core and have been extensively studied; indeed, combining various pairs of residues at the a and d positions of the heptad repeat induces specific oligomer state switches (Harbury *et al.*, 1993).

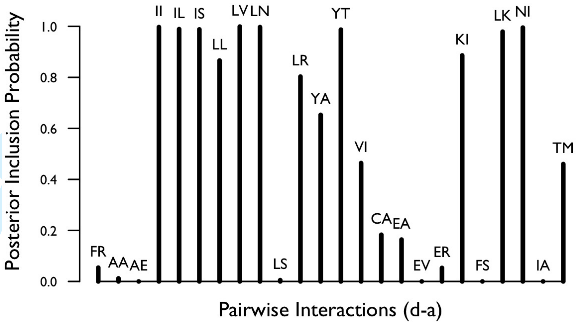

Fig. 2. Posterior inclusion probabilities for residues spaced 4 register positions apart, more specifically at the $d - a$ register positions. Residue pairs were preselected prior to being run through the OPS Bayesian variable selection model. Posterior inclusion probabilities were obtained after a run-length of 10000 iterations with 5000 burn-in iterations.

The posterior inclusion probabilities of pairwise associations at the $d - a$ pairs provided a noteworthy validation of the Bayesian variable selection scheme that was used. Indeed, the highest-scoring pairwise associations have been validated by experimental studies that incorporated known rules for hydrophobic cores into a designed background. For example, II (isoleucine at register d and a) is a well-established trimer-favouring pairwise interaction, as well as the LL (leucine at register d and a). LV (leucine at register d and valine at register a) associations is a dimer-favouring combination (Harbury *et al.*, 1993), while the IL (isoleucine at register d and leucine at register a) pairwise interaction has been shown to confer tetrameric conformations (Harbury *et al.*, 1993).

Because residues at the $d - a$ register pairs are placed in the hydrophobic core of coiled coils, it was expected that many pairwise associations would involve hydrophobic residues. However, a few high-scoring pairwise associations with polar residues were also found. Once again, these associations have been confirmed by experimental work that studied the effects of including a polar side chain within the otherwise hydrophobic core of a coiled-coil complex. For example, the LN (leucine at register d and asparagine at register a) or LK (leucine at register d and lysine at register a) pairs are found more often in dimeric sequences (Harbury *et al.*, 1993; Gonzalez *et al.*, 1996). Furthermore, the pairwise association NI (asparagine at register d and isoleucine at register a) or IS (isoleucine at register d and serine at register a) tends to favour trimeric structures (Hartmann *et al.*, 2009; Akey *et al.*, 2001).

The Bayesian variable selection scheme also detected pairwise associations, such as YT (tyrosine at register d and threonine at register a), that had no precedent in experimental studies. Some work has characterized the effects of placing the polar threonine residue at the a register position, which produce trimers (Akey *et al.*, 2001). However, this work is in the context of a clearly defined design background (*i.e.* leucine and isoleucine in the hydrophobic core positions) and does not consider the effects of simultaneously placing a tyrosine residue at the a register position. Placing tyrosine in the hydrophobic core may confer higher-order oligomerization, as only these structures could accommodate the large side-chain of tyrosine without disrupting the entire coiled-coil conformation (Walshaw and Woolfson, 2003). This hypothesis is supported by the PDB 2GUV structure, an engineered pentamer that incoporates bulky phenyalanine residues in its hydrophobic core (Liu *et al.*, 2006).

Overall, the posterior inclusion probabilities pairwise associations at the $d-a$ pairs, and obtained from first principles, agreed with experimental studies. Since the pairwise associations selected through Bayesian variable selection were coherent with experimental and structural interpretation, two main suggestions are proposed: (1) they represented valid pairwise associations to discriminate between coiled-coil oligomeric states, and should be included in the LOGICOIL predictive model; (2) they could be exploited to facilitate the rational design of coiled-coil structures. Here, we focussed on the first point, and investigated if the inclusion of pairwise residue effects at neighbouring positions of coiled-coil sequences improved the predictive power of LOGICOIL. This was achieved by applying Bayesian variable selection on the $i \rightarrow i+1$, $i \rightarrow i+3$ and $i \rightarrow i+4$ positions of the coiled-coil heptad repeat.

6.2 Performance of LOGICOIL

LOGICOIL was assessed using 10-fold cross validation on the pristine dataset, and variable selection was performed internally to the 10-fold cross validation scheme; *i.e.*, variables were re-selected whenever the training set and test set were changed. Prior to assessing the performance on a test fold, LOGICOIL learned from the training fold according to a three-step process:

1. Bayesian variable selection was run on each of the distinct $i \rightarrow i+1$, $i \rightarrow i+3$ and $i \rightarrow i+4$ spatial positions in the coiled-coil heptad repeat. For example, Bayesian variable selection was run on the seven distinct $i \rightarrow i+1$ positions $a-b, b-c, c-d, d-e, e-f, f-g$ and $g-a$. The inclusion probabilities of the pairwise associations were then computed from the output obtained for each spatial position.
2. The highest-scoring pairwise associations at all of the $i \rightarrow i+1$, $i \rightarrow i+3$ and $i \rightarrow i+4$ positions were identified, grouped together and ran through a secondary Bayesian variable selection scheme. Inclusion probabilities were computed from the final output and only the highest scoring pairwise associations were included in the LOGICOIL predictive model.
3. The parameter values of the retained predictors were fitted in a separate model, and the resulting MCMC samples were used to estimate parameters and class probabilities as in §3.3.

Generally, we observed that the selected pairwise associations remained uniform throughout the separate folds. The convergence behaviour of the secondary Bayesian variable selection scheme and the pairwise associations that were selected to be included in the LOGICOIL predictive model are listed in the §1 of the supplementary information. Figure 3 shows the average mAUC values and associated variations obtained by LOGICOIL during the 10-fold cross-validation scheme.

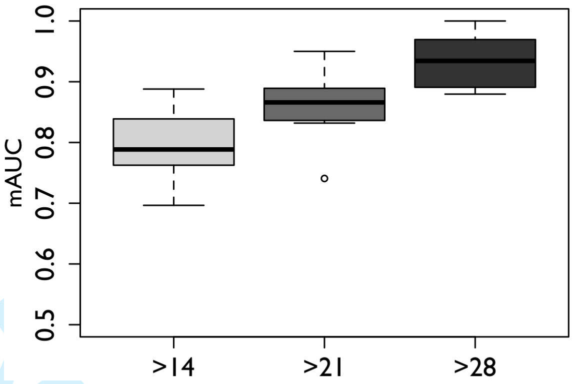

Fig. 3. Boxplot of the mAUC achieved by LOGICOIL when response probabilities were estimated using multinomial logistic regression. The mAUC were obtained using 10-fold cross-validation on coiled-coil structures with sequences greater in length than 14 residues (light grey), 21 residues (dark grey) and 28 residues (black) are shown.

The inclusion of pairwise residue effects drastically improved our ability to distinguish between coiled-coil sequences belonging to different oligomeric states. In particular, the performance obtained for coiled-coil sequences longer than 28 residues was high (mAUC = 0.93). As shown in Figure 4, pairwise AUC's of the different pair of classes were checked to ensure that the good overall performance was not a side product of a few, well-separated cases.

The separation rate for pairs involving tetramers was on average lower, in particular for coiled-coil sequences shorter than 28 residues. This could be attributed to the low numbers of tetramers in our training set, meaning that more structural information may be needed to improve predictions. While AUC measures are useful in estimating the discrimination accuracy of an algorithm, they do not explicitly report the fraction of correct assignments; *i.e.*, its accuracy. Therefore, we performed leave-one-out cross-validation on the LOGICOIL pristine dataset. We report the results as a confusion matrix obtained for coiled coils longer than 20 residues,

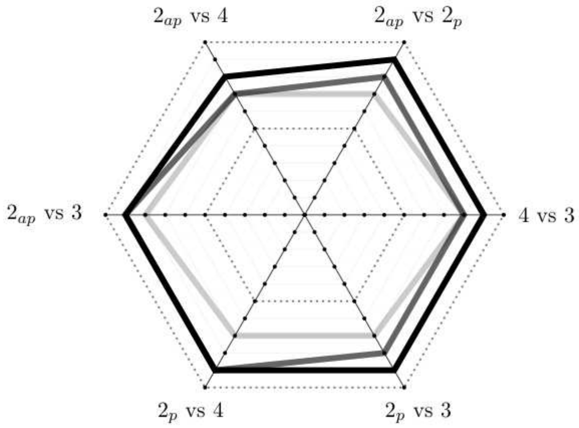

Fig. 4. Pairwise AUC values of the LOGICOIL predictive model for coiled coils with sequence length above14 (light grey), 21 (dark grey) and 28 (black) residues are shown. The symbol 2_{ap} represents antiparallel dimers, 2_p represents parallel dimers, 3 represents trimers and 4 represents tetramers.

shown in table 1, which gives correct predictions on the diagonal and incorrect predictions in the off-diagonal cells.

		Prediction Outcome				
		Antiparallel Dimers	Parallel Dimers	Trimers	Tetramers	Accuracy
Truth	Antiparallel Dimers	160	8	10	17	0.82
	Parallel Dimers	15	68	4	4	0.75
	Trimers	4	1	31	0	0.86
	Tetramers	6	0	1	22	0.76
						0.80

Table 1. Confusion matrix of the LOGICOIL predictions based on leave-one-out cross-validation of coiled-coil sequences in the pristine dataset with sequence length longer than 20 residues. The rows and columns show the true and predicted oligomeric states, respectively. The diagonal elements of the matrix show the number of correct assignments, while the off-diagonal shows the number of incorrect assignments.

On this basis, LOGICOIL showed high classification accuracies for predicting the oligomeric states of coiled-coil sequences in the pristine dataset. We observe that the highest misclassifications occurred between antiparallel and parallel dimers, which might be expected given the similarities between these structures (Walshaw and Woolfson, 2001a). Nevertheless, the increased predictive power of LOGICOIL suggested that spatial associations contributed to differentiate between parallel dimers, antiparallel dimers, trimers and tetramers. Interestingly, the separation rate achieved for the antiparallel dimer/tetramer pair was insensitive to changes in the coiled-coil sequence-length cutoff, thus raising questions on the inherent nature of the structures involved. Contrary to tetrameric coiled-coil structures, in-depth analysis of the antiparallel dimers structures in the LOGICOIL dataset suggested that the vast amount of available data for the antiparallel oligomer is in fact biased by many short, buried (*i.e.* packed within a protein structure) coiled coils. It is possible therefore that such structures are influenced into coiled coil-like conformation by interactions involved in tertiary protein structure. Hence, it could be argued that they are highly dependent on packing interactions in order to adopt a coiled-coil structure, and therefore do not represent ”real” — or free-standing — coiled coils. While it is possible that the performance of the LOGICOIL algorithm could be optimized if structures that exhibit this property were removed from its dataset, it was judged that this would have been too subjective. Additionally, selectively removing data from the training set could lead to unrepresentative database that would not reflect the known population of coiled coils observed in nature. In conclusion, although the caveats of the LOGICOIL dataset should be highlighted and taken into consideration, it is probably more prudent to strictly rely on the SOCKET-annotated data and not tamper with any of the dataset.

6.3 Comparing LOGICOIL with other algorithms

LOGICOIL is the only reported method capable of discriminating between multiple coiled-coil oligomeric states, which complicated the benchmarking of LOGICOIL against other existing coiled-coil oligomer-state predictors. Indeed, MultiCoil, Multicoil2, SCORER 2.0 and PrOCoil give predictions for parallel dimeric and parallel trimeric coiled coils only. Although it was already shown that LOGICOIL achieved high predictive accuracy, it was necessary to measure its performance relative to these other algorithms. To do this, LOGICOIL was reduced to a similar two-state predictor. The predictive powers of each algorithm were compared on the non-redundant dataset of parallel dimeric and trimeric coiled-coil sequences developed for the SCORER 2.0 algorithm (Armstrong *et al.*, 2011). To ensure fair comparison, LOGICOIL, SCORER 2.0 and PrOCoil were retrained and assessed using 10-fold cross-validation to provide independent tests of their utility. Due to the fact that MultiCoil and Multicoil2 could not be re-trained and can only score sequences longer than 21 residues, we only considered coiled coils with sequence length above this threshold.

The AUC values and classification accuracies displayed in Table 2 show that LOGICOIL outperformed all other algorithms in this two-state-prediction test, with PrOCoil coming in a close second. Due to the unavailability of the MultiCoil and Multicoil2 source code, it was not possible to carry out cross-validation for these algorithms. As a consequence, there is a possibility of overlap between the coiled-coil sequences contained in the test database and the training set of the concerned algorithms; thus biasing results. Nonetheless, MultiCoil and Multicoil2 did not compare favourably with other algorithms. The performance of MultiCoil was possibly linked to its training set, which is heavily biased towards long parallel two-stranded coiled coils (Gruber *et al.*, 2006). This caveat was resolved in the recently retrained and redesigned Multicoil2 predictor, which improved the performance of MultiCoil but was not to the level of the other predictors.

PrOCoil nearly matched LOGICOIL on two-state prediction, and it would be interesting to evaluate how this method performs on multi-state classification. SCORER 2.0 also achieved high discrimination rate, but is slightly surpassed by PrOCoil and LOGICOIL. Interestingly, SCORER 2.0 is the only algorithm to include no pairwise residue effects in its predictive model, which serves as a reminder that including pairwise residue effects does not necessarily increase predictive power. Thus, we suggest that pairwise-association information may not be so relevant in the context of two-state prediction, which is rationalized by

Algorithm	AUC	Recall Dimer / Trimer	Accuracy
LOGICOIL	0.90	1 / 0.88	0.95
SCORER 2.0	0.85	0.86 / 0.24	0.71
PrOCoil	0.89	0.97 / 0.60	0.88
MultiCoil	0.591	0.09 / 0.00	0.06
Multicoil2	0.66	0.35 / 0.20	0.32

Table 2. AUC values of LOGICOIL, SCORER 2.0, PrOCoil, MultiCoil and Multicoil2 when used to classify the oligomeric state of coiled coils on the pristine coiled-coil test set developed in (Armstrong *et al.*, 2011). Only coiled coils with sequence $>$ 20 amino acids were used, as MultiCoil and Multicoil2 do not accept any input shorter than 21 characters.

the well-characterized differences between parallel dimeric and trimeric coiled coils, *i.e*, there were enough distinctive features between the two structures so that the added sensitivity gained from spatial associations was not necessary. Rather, it is the construction of reliable training sets and methods to detect relevant pairwise associations that appears most important. For example, LOGICOIL and PrOCoil select pairwise associations with a scheme that favours sparsity. Also, their performance was reported with variable selection performed through external cross-validation, whereas there are no indication that the same procedure was applied during the original assessment of MultiCoil and Multicoil2. In addition, higher-order information may well be important in coiled-coil discrimination (Hadley *et al.*, 2008; Steinkruger *et al.*, 2010). We also assessed how straightforward homology-based searches using the BLAST algorithm performed on multi-state classification of coiled-coil oligomeric state prediction (see §2 in the supporting information). Again, the predictive accuracy obtained by LOGICOIL significantly outperformed the BLAST predictions.

7 CONCLUSIONS

This work has introduced LOGICOIL, the first algorithm to address the problem of predicting multiple coiled-coil oligomeric states from protein-sequence information alone. LOGICOIL increases our predictive coverage of the known coiled-coil structures from 31% to over 90%, but also distinctly improves our ability to differentiate between coiled-coil sequences of different oligomeric state. By taking into account the independent contribution of amino acids at different register positions, and subsequently including pairwise association effects between distantly positioned residues, we show that LOGICOIL achieves a high discrimination rate when predicting the oligomeric state of coiled-coil sequences across a range of structures, including antiparallel dimers, parallel dimers, trimers and tetramers. As the only algorithm allowing for such extensive coverage of the total coiled-coil population, LOGICOIL could not be benchmarked against any other algorithms. However, when constrained to the limitations of the currently available coiled-coil oligomeric state predictors, SCORER 2.0, MultiCoil and PrOCoil, it was demonstrated that LOGICOIL offers equal or superior predictive power. Furthermore, LOGICOIL

While the improvements that the LOGICOIL algorithm brings to the field of coiled-coil oligomeric state prediction are clear, we propose that LOGICOIL may still benefit from further iterations. Most notably, the LOGICOIL algorithm is currently constrained to the sole function of predicting coiled-coil oligomeric state, while depending on third-party softwares for the detection of coiled-coil domains in protein sequence. We suggest that extending the LOGICOIL predictive function to also include coiled-coil domain prediction would greatly benefit users, as this would result in a truly free-standing software capable of simultaneously predicting coiled-coil regions in a protein sequence along with its associated oligomeric state. Although LOGICOIL in its present form is capable of dealing with heptad breaks such as stutters, stammers and skips, it is not explicitly designed to account for the potential effects that may result from these sequence discontinuities. Given the increasing number of newly detected non-canonical coiled-coil sequences, future work will focus on the retrieval of heptad-break specific information to augment the predictive power and coverage of LOGICOIL. Despite these minor caveats, we suggest that the development of LOGICOIL, and the unique features it offers, widens the breadth of opportunities for biologists and bioinformaticians alike.

LOGICOIL is publicly and freely available via the world-wide web at the following URL: http://coiledcoils.chm.bris.ac.uk/LOGICOIL and can be used as stand-alone software for known coiled-coil regions, or in conjunction with MARCOIL, for coiled-coil region detection and oligomeric state assignment.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Craig Armstrong, Dr. Gail Bartlett, Dr. Drew thompson and members of the Woolfson lab for several useful discussions. The authors would also like to acknowledge Dr. Mauro Delorenzi for allowing the free use of MARCOIL on the LOGICOIL web server. We would also like to acknowledge Dr. Bodenhofer for providing useful information regarding the PrOCoil software.

Funding: TLV was funded by a BCCS PhD studentship from the EPSRC.

REFERENCES

Ai-Jun, Y. and Xin-Yuan, S. (2010). Bayesian variable selection for disease classification using gene expression data. *Bioinformatics*, **26**(2), 215–222.

Akey, D., Malashkevich, V. N., and Kim, P. S. (2001). Buried polar residues in coiled-coil interfaces. *Biochemistry*, **40**(21), 6352–6360.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, **88**(422), 669–679.

Armstrong, C. T., Vincent, T. L., Green, P. J., and Woolfson, D. N. (2011). SCORER 2.0: an algorithm for distinguishing parallel dimeric and trimeric coiled-coil sequences. *Bioinformatics, Advance Access*.

Becker, N., Werft, W., Toedt, G., Lichter, P., and Benner, A. (2009). penalizedSVM: a R-package for feature selection SVM classification. *Bioinformatics*, **25**(13), 1711–1712.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res*, **28**(1), 235–242.

Brown, P. J., Vanucci, M., and Fearn, T. (2002). Bayes model averaging with selection of regressors. *Journal of the Royal Statistical Society (B)*, **64**(3), 519–536.

Crick, F. H. C. (1953). The packing of α-helices - simple coiled coils. *Acta Crystallographica*, **6**(8-9), 689–697.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recogn Lett*, **27**(8), 861–874.

Fernandez, C., Ley, E., and Steel, M. F. J. (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, **100**, 381–427.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman and Hall, London, 2nd edition.

Gonzalez, L., Woolfson, D. N., and Alber, T. (1996). Buried polar residues and structural specificity in the GCN4 leucine-zipper. *Nat Struct Biol*, **3**, 1011–1018.

Green, P. J. (1995). Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**(4), 711–732.

Gruber, M., Söding, J., and Lupas, A. N. (2006). Comparative analysis of coiled-coil prediction methods. *J Struct Biol*, **155**(2), 140–145.

Gustafson, P. and Lefebvre, G. (2008). Bayesian multinomial regression with class-specific predictor selection. *The annals of Applied Statistics*, **2**(4), 1478–1502.

Hadley, E. B., Testa, O. D., Woolfson, D. N., and Gellman, S. H. (2008). Preferred side-chain costellations at antiparallel coiled-coil interfaces. *Proc. Natl. Acad. Sci. U. S. A.*, **105**, 530–535.

Hand, D. J. and Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problem. *Machine Learning*, **45**, 171–186.

Harbury, P. B., Zhang, T., Kim, P. S., and Alber, T. (1993). A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants. *Science*, **262**(5138), 1401–1407.

Hartmann, M. D., Ridderbusch, O., Zeth, K., Albrecht, R., Testa, O. D., Woolfson, D. N., Sauer, G., Dunin-Horkawicz, S., Lupas, A. N., and Alvarez, B. H. (2009). A coiled-coil motif that sequesters ions to the hydrophobic core. *Proc Natl Acad Sci USA*, **106**(40), 16950–16955.

Hochreiter, S. and Obermayer, K. (2006). Support vector machines for dyadic data. *Neural Computation*, **18**(6), 1471–1510.

Holmes, C. and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, **1**(1), 145–168.

Imai, K. and van Dyk, D. A. (2005a). A Bayesian analysis of the multinomial probit model using the marginal data augmentation. *Journal of Econometrics*, **124**(2), 311–334.

Imai, K. and van Dyk, D. A. (2005b). MNP: R package for fitting multinomial probit models. *Journal of Statistical Software*, **14**(3), 1–32.

Jones, D. T. and Swindells, M. B. (2002). Getting the most from PSI-BLAST. *TIBS*, **27**, 161–164.

Lamnisos, D., Griffin, J. E., and Steel, M. F. J. (2009). Transdimensional sampling algorithms for Bayesian variable selection in classification problems with many more variables than observations. *Journal of Computational and Graphical Statistics*, **18**, 592–612.

Lamnisos, D., Griffin, J. E., and Steel, M. F. J. (2010). Cross-validation prior choices in Bayesian probit regression with many covariates. forthcoming.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2007). CLUSTAL W and CLUSTAL X version 2.0. *Bioinformatics*, **23**(21), 2947.

Li, W. Z. and Godzik, A. (2006). CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**(13), 1658–1659.

Liang, F., Paulo, R., Molina, G., A, C. M., and O, B. J. (2008). Mixture of g-priors for Bayesian variable selection. *Journal of American Statistical Association*, **103**(410-423).

Liu, J., Zheng, Q., Deng, Y., Kallenbach, N. R., and Lu, M. (2006). Conformational transition between four and five-stranded phenylalanine zippers determined by a local packing interaction. *Journal of Molecular Biology*, **361**, 168–179.

Lupas, A. N. and Gruber, M. (2005). The structure of α-helical coiled coils. *Adv Protein Chem*, **70**, 37–78.

Mahrenholz, C. C., Abfalter, I. G., Bodenhofer, U., Volkmer, R., and Hochreiter, S. (2011). Complex networks govern coiled-coil oligomerization – predicting and profiling by means of a machine learning approach. *Molecular Cell Proteomics*, **10**(5), EPub.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall, London.

Moutevelis, E. and Woolfson, D. N. (2009). A periodic table of coiled-coil protein structures. *J Mol Biol*, **385**(3), 726–732.

O'Hara, R. B. and Sillanpaaa, M. J. (2009). A review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis*, **4**(1), 85–118.

Rackham, O. J. L., Madera, M., Armstrong, C. T., Vincent, T. L., Woolfson, D. N., and Gough, J. (2010). The evolution and structure prediction of coiled coils across all genomes. *J Mol Biol*, **403**, 480–493.

Sha, N., Vanucci, M., Tadesse, M. G., Brown, P. J., Dragoni, I., Davies, N., Roberts, T. C., Contestabile, A., Salmon, M., Buckley, C., and Falciani, F. (2004). Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. *Biometrics*, **60**, 812–819.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**(20), 3940–3941.

Steinkruger, J. D., Woolfson, D. N., and Gellman, S. H. (2010). Side-chain pairing preferences in the parallel coiled-coil dimer motif: Insight on ion pairing between core and flanking sites. *J Am Chem Soc*, **132**(22), 7586–7588.

Stingo, F. C. and Vanucci, M. (2010). Bayesian models for variable slection that incoporate biological information. *Bayesian Statistics*, **9**, 659–678.

Team, R. D. C. (1993). *R: A language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Testa, O. D., Moutevelis, E., and Woolfson, D. N. (2009). CC+: a relational database of coiled-coil structures. *Nucleic Acids Res*, **37**(Database issue), D315–22.

Trigg, J., Gutwin, K., Keating, A. E., and Berger, B. (2011). Multicoil2: predicting coiled coils and their oligomerization states from sequence in the twilight zone. *PLoS ONE*, **6**(8), e23519.

Tuchler, R. (2008). Bayesian variable selection for logistic models using auxiliary mixture sampling. *Journal of Computational and Graphical Statistics*, **17**, 76–94.

Walshaw, J. and Woolfson, D. N. (2001a). Open-and-shut cases in coiled-coil assembly: Alpha-sheets and alpha-cylinders. *Protein Science*, **10**, 668–673.

Walshaw, J. and Woolfson, D. N. (2001b). SOCKET: a program for identifying and analysing coiled-coil motifs within protein structures. *J Mol Biol*, **307**(5), 1427–1450.

Walshaw, J. and Woolfson, D. N. (2003). Extended knobs-into-holes packing in classical and complex coiled-coil assemblies. *Journal of Structural Biology*, **144**, 349–361.

Wolf, E., Kim, P. S., and Berger, B. (1997). MultiCoil: a program for predicting two-and three-stranded coiled coils. *Protein Sci*, **6**(6), 1179–1189.

Woolfson, D. N. and Alber, T. (1995). Predicting oligomerization states of coiled coils. *Protein Sci*, **4**(8), 1596–1607.

Yu, Y. B. (2002). Coiled-coils: stability, specificity, and drug delivery potential. *Advanced drug delivery reviews*, **54**(8), 1113–1129.

Zhou, X., Liu, K. Y., and Wong, S. T. C. (2004). Cancer classification and prediction using logistic regression with Bayesian gene selection. *Journal of Biomedical Informatics*, **37**, 249–259.

Zhou, X., Wang, X., and Dougherty, E. R. (2006). Multi-class cancer classification using multinomial probit regression with Bayesian gene selection. *IEEE Proceedings in Systems Biology*, **153**(2), 70–78.