

A Bayesian network model of proteins' association with promyelocytic leukemia (PML) nuclear bodies

Mikael Bodén^{1*}, Graham Dellaire², Kevin Burrage^{1,3}
Timothy L. Bailey¹

¹Institute for Molecular Bioscience,
The University of Queensland, QLD 4072, Australia

²Department of Pathology, Faculty of Medicine,
Dalhousie University, B3H 4H7, Canada

³Oxford Centre for Integrative Systems Biology,
University of Oxford, OX1 3QD, United Kingdom

June 23, 2009

Abstract

The modularity that nuclear organization brings has the potential to explain the function of aggregates of proteins and RNA. Promyelocytic leukemia nuclear bodies are implicated in important regulatory processes. To understand the complement of proteins associated with these intra-nuclear bodies, we construct a Bayesian network model that integrates sequence and protein-protein interaction data. The model predicts association with promyelocytic

*To whom correspondence should be addressed

leukemia nuclear bodies accurately when interaction data is available. At a false positive rate of 10%, the true positive rate is almost 50%, indicated by an independent nuclear proteome reference set. The model provides strong support for further expanding the protein complement with several important regulators and a richer functional repertoire. Using special SVM-nodes (equipped with string kernels), the Bayesian network is also able to produce predictions on the basis of sequence only, with an accuracy superior to that of baseline models.

1 Introduction

The structurally and functionally heterogeneous promyelocytic leukemia nuclear bodies (PML-NBs) sequester numerous important regulatory proteins (Bernardi and Pandolfi, 2007), and are critical to stress response and tumour suppression (Krieghoff-Henning and Hofmann, 2008). PML-NBs are intra-nuclear structures and in contrast to cytoplasmic organelles they are not membrane-bound. Instead association with PML-NBs is based on molecular interaction (Shen et al., 2006). Once imported into the nucleus, proteins are transported by diffusion and many proteins only associate transiently, and frequently with several other compartments (Gorski and Misteli, 2005). The list of proteins known to be co-localized with promyelocytic leukemia protein (Pml)—the signature protein of these structures—is growing. Although the PML-NB makes extensive contacts with surrounding chromatin—suggesting a possible role in DNA transactions such as transcription, replication or repair of DNA—the structure is largely devoid of RNA and DNA (Boisvert et al., 2000).

Sumoylation is a post-translational modification believed to be essential to many intra-nuclear processes. For instance, sumoylation is a key regulator in the assembly and disassembly of PML-NBs (Shen et al., 2006; Heun, 2007). The sumoylation process involves modifying a lysine by conjugating small ubiquitin-related modifier (Sumo) on the substrate. Many modified sites conform to the [LVI]KXE consensus pattern, and sumoylation may modulate the substrates' ability to interact with other proteins (Heun, 2007). Sumo can also

bind covalently to binding sites conforming to a [VI]X[VI][VI] consensus sequence. The Pml protein has several sumoylation sites and a binding site for covalent interactions with Sumo. When one or the other is inhibited, nuclear bodies are not formed (Shen et al., 2006).

Generally speaking, to predict system behaviour, diverse data sets are usefully combined with several other sources of knowledge and constraints. To shed light on the mechanisms used by non-membrane based microenvironments and to explore the implications for intra-nuclear protein organization, this paper aims to model the association of proteins with PML-NBs. Since association with such structures appears to largely be based on a multitude of molecular interactions, we propose a model that integrates sequence data and interaction networks. We develop powerful mechanisms for incorporating sequence prediction modules based on string kernels and support-vector machines into a probabilistic Bayesian network. These mechanisms enable the model to process entirely novel proteins for which interactions are not known. The model makes experimentally testable and quantitative predictions for the *M. musculus* nuclear proteome.

To model protein association with PML-NBs, it is natural to turn to interactions that involve known PML-NB members. Many functionally important gene products including Pml, Daxx, Blm, Hipk, Rad51, Tp53 and Crebbp localize to PML-NBs (Hofmann and Will, 2003). Table 1 provides a summary of interactions relevant to PML-NBs. A predominant view is that Pml may not bind directly to all PML-NB members, but instead establishes a supra-molecular scaffold necessary for their recruitment (Hofmann and Will, 2003).

There are at least seven known isoforms of Pml. Each isoform may have specific interactions that will affect PML-NB recruitment. Since interaction data seldom distinguishes between isoforms explicitly, we use isoform I of Pml as a canonical representative. The model presented herein thus cannot recognize that some PML-NB proteins are only known to co-localize with a specific Pml isoform.

Members of the nuclear body may share specific sequence features and domains (Bickmore and Sutherland, 2002), motivating the incorporation of such properties in a model's

Table
1 about
here.

decision. Two main prediction scenarios are evaluated in this paper. In the first, we assume that apart from its sequence, interactions of a query protein are known. In the second, the query protein is presented only by its sequence. We evaluate the expected accuracy of correct classification in both scenarios using the PML-NB association probability inferred by the model. The two scenarios differ only by having different inputs instantiated as described in the next section.

2 Methods and materials

Before describing the model that is used in this study, it is useful to consider the most basic approaches. Table 1 lists several candidates that a protein may use to associate with the nuclear body. However, in most cases only a fraction of their interactions involve another PML-NB protein. According to the data in Table 1, if a novel protein interacts with Pml, for example, the conditional probability of the novel protein associating with PML-NB is 0.32 (17/52). Additional interactions may be combined (probabilistically) to establish the association status with possibly greater accuracy. In such cases, we should also consider the prior probability of interactions (e.g. how often does a protein interact with Pml).

An alternative information source to interactions is the amino acid sequence of a protein. Support vector machines (SVMs) naturally assign class labels to amino acid sequences by using *string kernels*, chosen to compare sequences on basis of (for instance) shared k -mers (Spectrum kernel) or alignment (Local Alignment kernel). Hence, after training, a SVM can produce a score, for any novel sequence, that indicates its class. In the following we will present a method that combines the best of these two methods into a probabilistic modelling framework.

A model of PML-NB association

The model of PML-NB is based on a Bayesian network, in which nodes are stochastic (random) variables and directed edges (parent to child) represent causal relationships between such variables. A child node thus represents a conditional probability where the parents are the variables conditioned upon.

Our Bayesian network (see Figure 1) integrates features derived from two sources: amino acid sequences and protein interactions. The main classification variable, “AssocWithPmlNb”, is a Boolean variable that indicates whether a query protein is associated with PML-NBs or not. The network includes Boolean variables for protein interactions (set according to the query protein’s known interactions with key PML-NB members, here called “hubs”), and for associations with proteins in the sumoylation pathway (set according to the matching of two known consensus motifs).

To capture additional sequence features that help in assigning a class to a query protein, a support-vector machine (SVM) predicts their PML-NB membership (shown as “PredictAssocPmlNb” in Figure 1). The SVM uses a string kernel function that enables it to process amino acid sequence input. To accommodate the absence of protein interaction data, additional SVMs similarly predict for each hub and for each protein in our data set, whether they interact or not (the remaining leaf nodes in Figure 1). The output of each of these SVMs gives the (predicted) likelihood of the query protein interacting with the named hub.

From the model we can determine the probability of the classification variable (“AssocWithPmlNb”) given the interaction status for each of the hubs and the presence of possible sumoylation sites. Similarly, we can query the support for the same variable when only predictions from sequence are available, leaving the interaction parent nodes unspecified.

Bayesian network

A Bayesian network offers a theoretical framework in which probabilistic inference is made practical via the specification of dependencies between variables. In fact, the full joint prob-

Figure
1 about
here.

ability of all (instantiated) variables, x_1, x_2, \dots, x_N , can be computed from a simple product of the conditional probabilities made explicit in the network structure.

$$P(x_1, x_2, \dots, x_N) = \prod_{i=1}^N P(x_i | pa(X_i)) \quad (1)$$

where $pa(X_i)$ is the set of parent(s) of the i th variable. The meaning of an edge is thus that the child node’s distribution is conditionally independent of all other variables given the values of its parents.

Inference of $P(X|\mathbf{e})$ where X is the (uninstantiated) query variable, and \mathbf{e} is the available evidence, is based on the full joint probability. The procedure involves “summing out” the set of unobserved variables $\mathbf{y} \in \mathbf{Y}$,

$$P(X|\mathbf{e}) = \eta \sum_{\mathbf{y}} P(X, \mathbf{e}, \mathbf{y}) \quad (2)$$

where η is a normalizing constant (ensuring that probabilities of X ’s possible values add to 1).

Prior probabilities (represented by nodes without parents) are easily obtained from *training data* by relative counts of outcomes. Conditional probabilities (represented by nodes with parents) are similarly estimated from relative counts of outcomes but only from entries that fulfil the conditioning case of the parent nodes (in the discrete case forming a conditional probability table).

The example based on Table 1, where we noted that Pml interacts with 52 proteins of which 17 are PML-NB members, illustrates (in Bayesian network jargon) a child variable (“AssociatesWithPmlNb”) that has one parent (“InteractWithPml”). It has two entries, one for when interaction with Pml is true, one for when it is false. The former will be assigned the probability 0.32, the latter is not disclosed in the table but can be similarly established (by counting the total number of samples that has Pml-interaction set to false, and the number of samples when it is set to false and the association with PML-NB is set to true). The prior

for the parent is simply the ratio of proteins that interact with Pml over the total number of proteins.

When the number of parent variables is large, the combinatorial number of possible conditioning cases make learning conditional probabilities problematic. With N Boolean parents, there are 2^N entries in the conditional probability table, each of which has a probability that is based on only those samples that exactly match the condition. This can be overcome at the cost of limiting the types of distributions we can model. Specifically, we restrict a child variable to be determined by the *Boolean Noisy-OR function*.

In contrast to a deterministic OR-gate, a Boolean Noisy-OR table allows for uncertainty of each parent variable to cause the child to be true (Diez, 1993). It makes the simplifying assumption that the inhibition of each parent variable (the extent by which the child is not made true) is independent of any other parents. Thus, the whole table of an instantiated variable x_i can be specified with as many probabilities as there are parent nodes, and the entire table can be filled in using the equation

$$P(x_i|pa(X_i)) = 1 - \prod_{X_j \in pa(X_i)} (1 - P(x_i|X_j)). \quad (3)$$

When no parents are true (T), the Noisy-OR conditional probability is 0.

We use a standard Noisy-OR table to alleviate issues of populating the seven-parent table for “AssocWithPmlNb”, reducing the number of probabilities to be estimated from the data sets from $2^7 = 128$ to 7.

Kernels and probabilistic SVMs

Through the use of kernels, an SVM can be trained to classify many kinds of data, including amino acid sequences. Not only do we wish to include predictions, but the ability of kernels to process data that is not naturally perceived as random variables motivates the incorporation of SVMs into the Bayesian framework. However, the SVM produces a prediction that is not a probability but a score (typically thresholded at 0 to discriminate between a positive and a

negative class),

$$f(\mathbf{x}) = \sum_{i=1}^M y_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) + b, \quad (4)$$

where \mathbf{x} is the query input, \mathbf{x}_i is the i^{th} training sample, $y_i \in \{+1, -1\}$ is the target class of sample i , α_i is the i^{th} Lagrange multiplier and b is a bias, each of which is tuned during training (Schölkopf and Smola, 2002). If α_i is non-zero, then \mathbf{x}_i is called a support vector. k is the kernel function that encapsulates an inner product between the two data points. String kernels assume that inputs are on the form of sequences of arbitrary length with symbols drawn from a finite set. They have been used extensively with SVMs to predict biologically meaningful features from both DNA and amino acid sequences. We test a number of different string kernels, including the spectrum kernel ($l \in \{1, 2, 3\}$), the mismatch and wildcard kernels ($l = 3$ and $m = 1$) (Leslie and Kuang, 2004), and the local alignment kernel ($\beta = 0.01$) (Saigo et al., 2004).

Similar to Hastie and Tibshirani (1996), we introduce the SVM as a continuous random variable f conditioned on a class variable, $p(f|y = +1)$ and $p(f|y = -1)$. Specifically, we fit two Gaussian densities on scores produced by samples from the positive ($y = +1$) and the negative class ($y = -1$), respectively. Unlike Hastie and Tibshirani (who go on to define a posterior class probability by constraining the densities) we estimate both classes' means and variances. The training data is divided into 3/4 of the set for training f and 1/4 for training p . In our Bayesian network, the class is the parent node, and the posterior class probability is thus given by Bayes' rule (an example is provided in Figure 2). It is also possible to incorporate SVMs into a probabilistic framework by using a parametric model, say a sigmoid function, to fit the class posterior directly (Platt, 2000). Barutcuoglu et al. (2006) use similar SVM-based nodes to predict individual functional features of genomic data and then use a Bayesian network to constrain the predictions to agree with a taxonomy such as the Gene Ontology.

SVMs use regularization parameters (the so-called C -values that constrain the α s in

Equation 4). Since the data is highly unbalanced (many more negatives than positives), we performed preliminary trials to determine settings used in all simulations. C is 10000 and 0.1 for the positive and negative class, respectively. Since the proportion of positives is similar in each classification problem, these values are used for both predicting the main association class and all interaction classes.

Simulation methodology

We only report on results using data not seen during training. All results are generated by 10-fold cross-validation. That is, we divide the data into ten subsets, and use nine for training and one for testing. The training and testing procedure is repeated, trying all ten combinations of the subsets. Consequently, each protein will be a test case in exactly one model.

To evaluate our models we first note that, due to the high number of negatives in the data sets, a percentage accuracy is not adequate. Instead we rely on the area under the ROC curve (AUC), which looks at the accuracy for all possible settings of sensitivity/specificity trade-offs. An AUC of 1.0 corresponds perfect classification accuracy and 0.5 indicates an accuracy that is no better than chance. The true positive rate (the y -axis in the ROC curve) is standardly defined as the ratio between the number of correctly predicted positives and the total number of positives. Conversely, the false positive rate (the x -axis in the ROC curve) is the ratio between the number of incorrectly predicted negatives and the total number of negatives.

Data sets

Recently the mouse nuclear proteome was defined to contain 2568 proteins with direct evidence of localization (Nucprot; Fink et al., 2008). This number almost doubles when including confidently predicted proteins. We identified a set of mouse PML-NB proteins by selecting the mouse orthologs in Nucprot of proteins annotated in the Nuclear Protein Database

(NPD; Dellaire et al., 2003) with the term “PML bodies”. We call this positive set of 76 proteins “Pos-all”.

It needs to be emphasized that association with intra-nuclear compartments in the NPD is established via a plethora of different experimental techniques, including gene-trap screens, mass spectrometry, and fluorescence tagging. Proteins that are not flagged as PML-NB proteins in NPD, are here, in the absence of evidence to the contrary, considered negatives. To increase the quality of the set, we first exclude all negatives that do not have experimental evidence of nuclear localization (i.e. not part of the 2568-set) and all proteins that are “un-reviewed” in Uniprot. We identified all negatives that interact with at least one PML-NB protein (we call this set “Neg-interact”). All negatives without any interactions with PML-NB proteins were subject to a filtering step requiring them to exhibit less than 10% sequence similarity to remove redundant entries (103 entries).

Some kernel functions are computationally expensive, making the large number of negatives difficult to deal with. As a practical compromise, the 1431 remaining non-interacting proteins were put into two groups “Neg-1” and “Neg-2”. Two sets are used in the simulations below: “All-1” (the union of “Pos-all”, “Neg-interact” and “Neg-1”; 1312 proteins) and “All-2” (the union of “Pos-all”, “Neg-interact” and “Neg-2”; 1313 proteins). The data sets are summarized in Table 2.

Due to the sparsity of localization data available it is very likely that many proteins designated as negatives are in fact associated with PML-NBs. (Indeed, as discussed later, several discoveries reported in the recent literature are not annotated in the NPD with “PML bodies” and are thus not included in our positive set.)

Several high-throughput experimental technologies have contributed to the construction of large-scale protein-protein interaction networks. To increase coverage we use (the union of) both the Molecular Interaction database (Chatr-aryamontri et al., 2007) and the Human Protein Reference Database (Mishra et al., 2006) interaction networks. By first finding mouse orthologs, any pair of proteins in our data sets can be queried for its interaction status (i.e. true or false). It is noted that such data is unreliable but that broad support for inferring

Table
2 about
here.

intra-nuclear association can be sought from it. Furthermore, it is desirable for a model not to rely on such data but be able to utilize it when available. There are several examples in the literature (Ben-Hur and Noble, 2005; Qiu and Noble, 2008) where sequence data is used (in full or complemented by other data) to predict whether a pair of proteins interacts or not, justifying our attempt to use this information to influence the model's decision.

3 Results

Intuitively, the protein interaction data should provide good indication of compartment association. From estimating the probability of PML-NB association given whether a novel protein interacts with Pml (as suggested earlier), the resulting AUC on "All-1" is a modest 0.60. Using Pml this way renders the best AUC of all the proteins in Table 1. If we group five proteins, and assume that the interaction a protein has with them is independent, the highest AUC we see is 0.67. Due to lack of data, a standard conditional probability table does not work. However, if we instead use a Noisy-OR conditional probability table (the Bayesian network in Figure 1 *without* the sumoylation-related and the prediction nodes) the AUC increases to 0.70. By exploring each protein's interaction with different proteins in Table 1 this way, we found that the 5-hub combination Pml, Sumo1, Blm, Rad51 and Mdm2 generally produces robust and accurate predictions of PML-NB association. In the following tests, these five hubs are used.

In the Bayesian network (see Figure 1), we set the conditional probabilities needed for inference by inspecting lists of PML-NB members, and lists of pairwise protein interactions as follows. For each protein in these sets, PML-NB association status, amino acid sequence and interaction status with all other proteins are known and are used for training and testing the model.

To present a protein to the Bayesian network, we first check whether it interacts with any of the hubs. The corresponding hub node is thus set to true or false, assuming that all interactions are available during training. (During testing a hub node can also be left

unspecified as described later.) Table 3 gives a typical Noisy-OR conditional probability table for “AssocWithPmlNb”.

Table
3 about
here.

Nodes fitted with an SVM (bottom row in Figure 1) are trained on the appropriate sequence data. “PredictAssocPmlNb” scores PML-NB association status. “PredictIntact-” nodes are trained on the same sequences, but with a different target: to predict interaction with the designated hub.

In more detail, the node “PredictAssocPmlNb” uses an SVM that takes the sequence as input and produces a score that influences—using standard Bayesian inference via two class likelihoods—the posterior belief in “AssocWithPmlNb”. In preliminary tests we observed that the so-called local alignment kernel (Saigo et al., 2004) provides consistently high classification accuracy when used inside standalone SVMs. Hence, the “PredictAssocPmlNb” is fitted with this kernel.

We use only about 3/4 of the sequence training data to train the SVM parameters. The densities, representing the probability of an SVM score given the class of sequence, are estimated from the scores for the remaining 1/4 of the training data. A typical pair of densities is shown in Figure 2. There is a substantial overlap between the score of the positives and the negatives but it is clear that this sequence scoring variable carries information that can be used by the parent node.

Figure
2 about
here.

Due to the semantics of Bayesian networks, the “PredictIntact-” nodes only influence inference when their parent is unspecified (i.e. with the “IntactWith-” variable is uninstantiated). Several test scenarios (described later) explore the impact of leaving some variables unspecified.

Tests using interactions

We perform standard exact inference to perform predictions of the “AssocWithPmlNb” variable using the test set as evidence (see Equation 2). In all tests, this probability is used directly as the model’s prediction of the query associating with PML-NBs.

Using the same five interaction hubs as above, the average AUC is 0.74 (on “All-1”). The test was repeated 10 times with different data set splits and the standard deviation was less than 0.01 (0.006). We also repeated the simulation with different combinations of interaction hubs. In some cases we noted even higher accuracies (0.76) but treat such conservatively, mindful of selection biases.

The predictions with high probability are usually quite reliable. In real terms, if we accept a false positive rate of 10%, the true positive rate is almost 50% (this corresponds to a probability threshold of 0.25). The specificity of the predictions drops at lower probabilities. If we want a true positive rate of 80% we need to accept a false positive rate of nearly 50%. (When interpreting these numbers one needs to consider that the negative set is much larger than the positive set. A large negative set automatically implies a low false positive rate since this is the ratio between incorrect positive predictions and the greater total of negative samples.) On “All-2” the average AUC is also 0.74. The fact that the accuracy is essentially unchanged between the two scenarios (trained and tested with different data sets) illustrates that the model’s performance is robust.

Novel predictions

The preceding tests show that PML-NB association can be predicted with good accuracy at least when protein interaction data is available for the query protein. We collected all predictions for proteins in “All-1” and “All-2”, when withheld from training.

Among the 20 predictions with a probability exceeding 0.65, 10 are annotated as PML-NB proteins in our reference set. The other half of the top-scoring predictions are proteins that are not currently annotated in the NPD to localize with PML-NBs (shown in Table 4). However, since the proteome is only sparsely annotated with intra-nuclear localization in general, and that the the reference set uses negatives assumed from absence of annotation to the contrary, we searched the literature to explore the possibility that the assumed negatives assigned high probabilities indeed associate with PML-NBs. Re-assuringly, we found

evidence of PML-NB localization for several proteins clearly mis-annotated as “negatives”.

The top prediction Wrn is known to cause Werner syndrome and is involved in DNA damage repair—a core PML-NB activity. Indeed, Wrn has been reported to be a body component under DNA damage stress (Blander et al., 2002). Rb1 (at probability 0.70) co-localizes with Pml within nuclear bodies (Alcalay et al., 1998). Sumo2 (0.70) localizes to PML-NBs (but also to nucleoli) (Fu et al., 2005). Further down the list of predictions, there is evidence that Pias2 (at probability 0.54) is partially localized to PML nuclear bodies (Tussie-Luna et al., 2002). Finally, after heatshock, a subset of PML-NBs co-localize with Hsf1 (0.50) (Hong et al., 2001). In total, 53 predictions have probabilities exceeding 0.5. We remove known positives (including the aforementioned examples of mis-annotations) and list the 27 remaining proteins in the Supplementary Material.

Table
4 about
here.

Predicting the role of PML-NBs

This section draws on the model to investigate the biological functions in which PML-NBs may be involved. Specifically, we check the statistical enrichment of Gene Ontology (GO) terms in the *predicted* set. Our null hypothesis is that, if proteins are sorted according to their PML-NB association probability, proteins with a specific term T appear in random order. The one-tailed Wilcoxon Ranksum test determines whether we can reject the null hypothesis and state that T is significantly associated with high-scoring proteins.

As a baseline, we use the Fisher Exact test to establish enrichment of specific terms in the reference data set (with 76 definitive PML-NB members). Here there are, for each term (T), and each protein, four possibilities: a protein is a PML-NB member or not, a protein is assigned a term T or not. The one-tailed test can reject a null hypothesis that states that the assignment of a term T is not biased by our knowledge of the PML-NB membership. In both statistical tests, we need to correct for multiple tests (all terms assigned to *any* protein in the nuclear proteome are looked at).

As before, we repeat simulations, training on both “All-1” and “All-2”. In each case

we identified over-represented terms, noting that almost always the same GO terms showed up. Using a model based on “All-1” (for training) as a representative, we retrieved 49 over-represented GO terms with an E -value less than 1. That is, using the p -value provided by the test, and the number of tests, we would expect less than 1 predicted positive by chance (a “Bonferroni correction”). With the exception of one GO term (GO:0044427 “Chromosomal part”), this list is a super-set of the list we get by using the PML-NB protein vs. non-PML-NB protein training data ($E < 0.05$). The list is provided in Table 5, where we also note those GO terms that are significantly associated with the set of known PML-NB members ($E < 1$). Notable additions only available from the model’s predictions (and not from the two-class training data) include the GO terms “SUMO ligase activity”, “protein sumoylation”, “kinase activity”, “phosphorylation” and “nucleotide binding”. The list also supports the belief that PML-NB plays important regulatory roles in, for instance, “induction of apoptosis (from intracellular signals)” (Hofmann and Will, 2003).

Table
5 about
here.

Predicting from sequence

The usefulness of prediction increases if we can operate solely on sequence. For example, the model can then map complete genomes. What can we say about PML-NB association of a protein from its sequence alone? In other words, what accuracy can we expect when the model is using only *predicted* protein-protein interactions?

The probability of interaction with hubs on basis of sequence is encoded into the Bayesian network by the child nodes of each hub. More specifically, the child node of a hub is fitted with an SVM that produces a score for the sequence indicating the SVM’s belief of it interacting with the hub. Predicting interaction from sequence is inherently difficult and we have limited numbers of positives for the hub proteins. To find suitable kernel functions for the child nodes, we first surveyed the accuracy of SVMs predicting the interaction status for each hub separately. The best kernel functions were then used in each prediction node of the Bayesian network. We list the kernel function that renders the best interaction prediction

AUC values for each of the five hubs in Table 6.

Sequence-only prediction of PML-NB association is done by setting the child nodes in accordance with the query protein. The two parent nodes “HasSumoylSite” and “HasSIM” are always set by using sequence data (by matching of motifs). The interaction parent nodes are left unspecified and are only inferred from the respective child nodes. Again we determine the model’s ability to predict “AssocWithPmlNb” in accordance with our data sets.

Using the same five hubs as evaluated above, both “All-1” and “All-2” render the same cross-validated average AUC of 0.64 ± 0.01 . However, there were many other combinations that gave better accuracy. For example, using the proteins Pml, Sumo1, Pias4, Tp53 and Crebbp as hubs, the AUC is 0.67. Interactions with these proteins are easier to predict from sequence than some of the others. We conservatively stay with the original five hubs, as this model is not chosen on basis of its separate ability to predict pairwise interaction features.

For comparison, simpler “baseline” models—for example, an SVM trained to classify each protein from the sequence—can be envisaged. This is exactly what the node “PredictAssocPmlNb” does. We note that by exploring the same set of kernel functions and the same density estimation procedure, this node alone can be used to predict PML-NB association (the class posterior probability) with an AUC of 0.62 ± 0.03 . As noted previously, the best kernel for this node is the local alignment kernel. This SVM can thus be seen as providing a baseline for predicting association with PML-NBs, inferior to the Bayesian network model. We also tried a standard SVM fitted with a logistic function as prescribed by (Platt, 2000), again with AUC 0.62 ± 0.03 . The variance in accuracy when relying only on sequence data, is thus consistently higher than the model that incorporates interaction data. Anecdotally, we notice that even when interaction data is included, candidates with strong homology to other proteins in the data set, e.g. Sumo2 and Prmt8 in Table 4, receive widely changing support by models between cross-validation runs.

We summarize the accuracy of the Bayesian network-based model and the baseline models in Table 7. An advantage of our Bayesian network model is that it can benefit from incomplete protein interaction data. In Table 7 we also provide the AUC rendered by the model

Table
6 about
here.

when protein interactions with a *single* hub are assigned for each query. The remaining interactions are predicted as in the tests that have access to sequence only. In most cases, the accuracy goes up slightly with the addition of known interaction data. Surprisingly, we note that by supplying only interactions with Rad51 and with all other interactions predicted, the overall accuracy drops compared to when the model only uses predicted interactions.

In summary, the model herein provides a small improvement when only sequence is considered (ten different data set splits for cross-validation render $p < 0.05$ using a paired t -test on the AUC for the Bayesian network being equal to that of the standard SVM). The model provides a substantial boost in accuracy when knowledge of protein interaction is considered, but most importantly it allows further inspection and validation of the model.

Table
7 about
here.

4 Discussion

The current model does not represent co-expression of interacting components. Proteins such as Blm and Wrn are associated with PML-NBs under certain phases of the cell cycle. Moreover, some proteins are only localized with PML-NBs under cell stress (e.g. DNA damage), in specific cell types, or when post-translationally modified. Predictions should thus be interpreted with this important caveat in mind. We expect that co-expression can be incorporated in a fashion analogous to the paired interactions. That is, via nodes that are set true for a query protein when it is co-expressed with specified key PML-NB member.

The gene ontology analysis highlighted the prevalence of sumoylation and phosphorylation amongst predicted members. Sumoylation is mechanistically linked to PML-NB localization (Shen et al., 2006), and Rb1 is one example that associates with PML-NBs in a phosphorylated state. We thus expect that extending the model to encompass modifications beyond sumoylation will increase its accuracy.

Bayesian networks have several advantages over other prediction methods. In stark contrast to a conventional “black-box” predictive model, a Bayesian network provides us with explanations about the domain and identifies important factors for specific predictions.

The natural integration of knowledge of PML-NB formation (based on protein interaction and sumoylation) into the Bayesian network plays an important role in improving the prediction accuracy. The Bayesian network requires only access to the sequence of a protein and then provides a minor but statistically significant improvement compared to our best support-vector machine. The novel incorporation of nodes that take sequence data as input dramatically increases the versatility of Bayesian networks.

When queried using complete or partial interaction data, the Bayesian network achieves an accuracy far exceeding that of our baseline SVM. More importantly, the model provides real biological insights. Our Bayesian network shows that Pml, Sumo1, Blm, Rad51 and Mdm2 form an effective group of key interaction hubs, assisting the accurate prediction of PML-NB membership.

Using our best model we predict several novel PML-NB members. Our top-predictions even re-discover several members known in the literature that were mistakenly labelled as negatives in our data set. (Incorrectly labelled proteins also impact negatively on our estimated accuracies, to err on the conservative side.) The model can assist in identifying the full PML-NB proteome experimentally. We use the model to identify a range of gene ontology terms that characterize the functions carried out by proteins linked to PML-NBs. We note important regulatory roles not discovered using statistical analysis from our “gold standard” positive set, including apoptosis and programmed cell death, phosphorylation, protein sumoylation, and SUMO ligase activity.

The flexible framework allows the current model to be incorporated into a larger model, representing both other intra-nuclear compartments and additional factors (like protein turnover rates, post-translational modifications and expression levels) to allow the fluid exchange to be recognized.

Acknowledgement

The authors thank Nurul Mohamad for preparing data sets and Dr Rohan Teasdale for helpful advice.

References

- Alcalay, M., Tomassoni, L., Colombo, E., Stoldt, S., Grignani, F., Fagioli, M., Szekeley, L., Helin, K., and Pelicci, P. G. (1998). The promyelocytic leukemia gene product (PML) forms stable complexes with the retinoblastoma protein. *Mol Cell Biol*, 18(2):1084–1093.
- Ariumi, Y., Ego, T., Kaida, A., Matsumoto, M., Pandolfi, P. P., and Shimotohno, K. (2003). Distinct nuclear body components, PML and SMRT, regulate the trans-acting function of HTLV-1 Tax oncoprotein. *Oncogene*, 22(11):1611–1619.
- Barutcuoglu, Z., Schapire, R. E., and Troyanskaya, O. G. (2006). Hierarchical multi-label prediction of gene function. *Bioinformatics*, 22(7):830–836.
- Ben-Hur, A. and Noble, W. S. (2005). Kernel methods for predicting protein-protein interactions. *Bioinformatics*, 21(S1):i38–i46.
- Bernardi, R. and Pandolfi, P. P. (2007). Structure, dynamics and functions of promyelocytic leukaemia nuclear bodies. *Nat Rev Mol Cell Biol*, 8(12):1006–1016.
- Bickmore, W. and Sutherland, H. (2002). Addressing protein localization within the nucleus. *The EMBO Journal*, 21:1248–1254.
- Blander, G., Zalle, N., Daniely, Y., Taplick, J., Gray, M. D., and Oren, M. (2002). DNA damage-induced translocation of the Werner helicase is regulated by acetylation. *J Biol Chem*, 277(52):50934–50940.
- Boisvert, F. M., Hendzel, M. J., and Bazett-Jones, D. P. (2000). Promyelocytic leukemia

- (PML) nuclear bodies are protein structures that do not accumulate RNA. *J Cell Biol*, 148(2):283–292.
- Chatr-aryamontri, A., Ceol, A., Palazzi, L. M., Nardelli, G., Schneider, M. V., Castagnoli, L., and Cesareni, G. (2007). MINT: the Molecular INTeraction database. *Nucleic Acids Res*, 35(Database issue):D572–D574.
- Conlan, L. A., McNeese, C. J., and Heierhorst, J. (2004). Proteasome-dependent dispersal of PML nuclear bodies in response to alkylating DNA damage. *Oncogene*, 23(1):307–310.
- Dellaire, G., Farrall, R., and Bickmore, W. (2003). The Nuclear Protein Database (NPD): sub-nuclear localisation and functional annotation of the nuclear proteome. *Nucl. Acids Res.*, 31(1):328–330.
- Diez, F. J. (1993). Parameter adjustment in Bayes networks: the generalized noisy or-gate. In *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence*, pages 99–105.
- Fink, J. L., Karunaratne, S., Mittal, A., Gardiner, D., Hamilton, N., Mahony, D., Kai, C., Suzuki, H., Hayashizaki, Y., and Teasdale, R. (2008). Towards defining the nuclear proteome. *Genome Biol*, 9(1):R15.
- Fu, C., Ahmed, K., Ding, H., Ding, X., Lan, J., Yang, Z., Miao, Y., Zhu, Y., Shi, Y., Zhu, J., Huang, H., and Yao, X. (2005). Stabilization of pml nuclear localization by conjugation and oligomerization of sumo-3. *Oncogene*, 24(35):5401–5413.
- Gorski, S. and Misteli, T. (2005). Systems biology in the cell nucleus. *J Cell Sci*, 118(Pt 18):4083–4092.
- Hastie, T. and Tibshirani, R. (1996). Classification by pairwise coupling. Technical report, Stanford University and University of Toronto.
- Heun, P. (2007). Sumorganization of the nucleus. *Curr Opin Cell Biol*, 19(3):350–355.

- Hofmann, T. G. and Will, H. (2003). Body language: the function of PML nuclear bodies in apoptosis regulation. *Cell Death Differ*, 10(12):1290–1299.
- Hong, Y., Dagger, Rogers, R., Matunis, M. J., Mayhew, C. N., Goodson, M., Park-Sarge, O.-K., and Sarge, K. D. (2001). Regulation of heat shock transcription factor 1 by stress-induced SUMO-1 modification. *J. Biol. Chem.*, 276:40263–40267.
- Krieghoff-Henning, E. and Hofmann, T. G. (2008). Role of nuclear bodies in apoptosis signalling. *Biochim Biophys Acta*, 1783(11):2185–2194.
- Leslie, C. and Kuang, R. (2004). Fast string kernels using inexact matching for protein sequences. *Journal of Machine Learning Research*, 5:1435–1455.
- Mishra, G. R., Suresh, M., Kumaran, K., Kannabiran, N., Suresh, S., Bala, P., Shivakumar, K., Anuradha, N., Reddy, R., Raghavan, T. M., Menon, S., Hanumanthu, G., Gupta, M., Upendran, S., Gupta, S., Mahesh, M., Jacob, B., Mathew, P., Chatterjee, P., Arun, K. S., Sharma, S., Chandrika, K. N., Deshpande, N., Palvankar, K., Raghavnath, R., Krishnakanth, R., Karathia, H., Rekha, B., Nayak, R., Vishnupriya, G., Kumar, H. G. M., Nagini, M., Kumar, G. S. S., Jose, R., Deepthi, P., Mohan, S. S., Gandhi, T. K. B., Harsha, H. C., Deshpande, K. S., Sarker, M., Prasad, T. S. K., and Pandey, A. (2006). Human protein reference database–2006 update. *Nucleic Acids Res*, 34(Database issue):D411–D414.
- Misteli, T. (2007). Beyond the sequence: cellular organization of genome function. *Cell*, 128(4):787–800.
- Mladenov, E., Anachkova, B., and Tsaneva, I. (2006). Sub-nuclear localization of Rad51 in response to DNA damage. *Genes Cells*, 11(5):513–524.
- Pal, S. and Sif, S. (2007). Interplay between chromatin remodelers and protein arginine methyltransferases. *J Cell Physiol*, 213(2):306–315.
- Platt, J. (2000). Probabilistic outputs for support vector machines and comparison to regu-

- larized likelihood methods. In Smola, A., Bartlett, P., Schölkopf, B., and Schuurmans, D., editors, *Advances in Large Margin Classifiers*. MIT Press, Cambridge, MA.
- Qiu, J. and Noble, W. S. (2008). Predicting co-complexed protein pairs from heterogeneous data. *PLoS Comput Biol*, 4(4):e1000054.
- Saigo, H., Vert, J.-P., Ueda, N., and Akutsu, T. (2004). Protein homology detection using string alignment kernels. *Bioinformatics*, 20(11):1682–1689.
- Schölkopf, B. and Smola, A. (2002). *Learning with kernels*. MIT Press, Cambridge, MA.
- Shen, T. H., Lin, H.-K., Scaglioni, P. P., Yung, T. M., and Pandolfi, P. P. (2006). The mechanisms of PML-nuclear body formation. *Mol Cell*, 24(3):331–339.
- Tussie-Luna, M. I., Michel, B., Hakre, S., and Roy, A. L. (2002). The SUMO ubiquitin-protein isopeptide ligase family member Miz1/PIAS β /Siz2 is a transcriptional co-factor for TFII-I. *J Biol Chem*, 277(45):43185–43193.

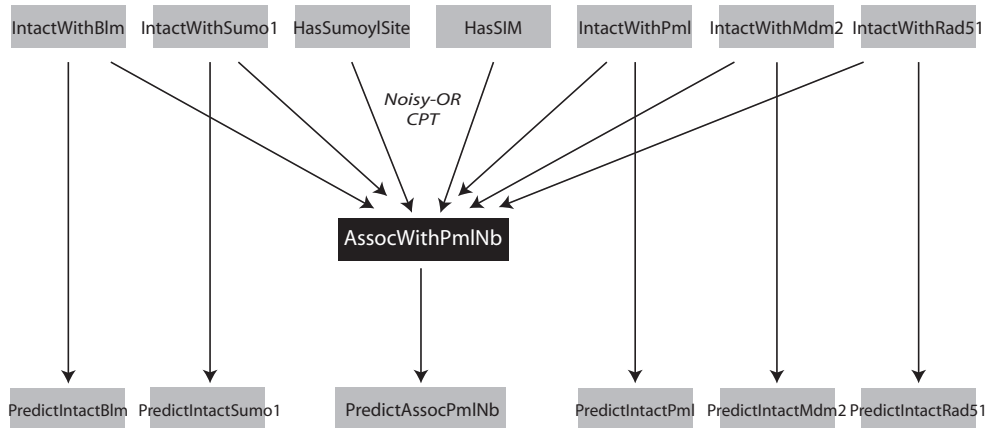


Figure 1: **The Bayesian network.** The main random (Boolean) variable of interest for prediction is “AssocWithPmlNb”. It represents the probability of a query protein being associated with PML-NBs. A range of variables modulate the belief, including protein interactions and the presence of possible sumoylation sites (top nodes; all Boolean). These variables are instantiated when such data is available. “IntactWithBlm” is true when the query protein interacts with Blm, false if it does not. All the other “IntactWith”-variables are set analogously. “HasSumoylSite” is true when the protein has a match to the sumoylation consensus motif anywhere along its sequence, and false otherwise. “HasSIM” is set analogously but uses the Sumo interaction consensus motif (SIM). PML-NB association and protein interactions can be predicted using nodes fitted with SVMs (bottom nodes; all Boolean). From the continuous SVM score, two class likelihood densities are estimated (one for samples that are “positives”, one for samples that are “negatives”). “PredictAssocPmlNb” is the score produced by the SVM that is trained to discriminate between PML-NB members and non-members. The “PredictIntact”-variables represent the scores for the SVMs that discriminate between hub binders and non-binders (where the hub is identified by the variable name).

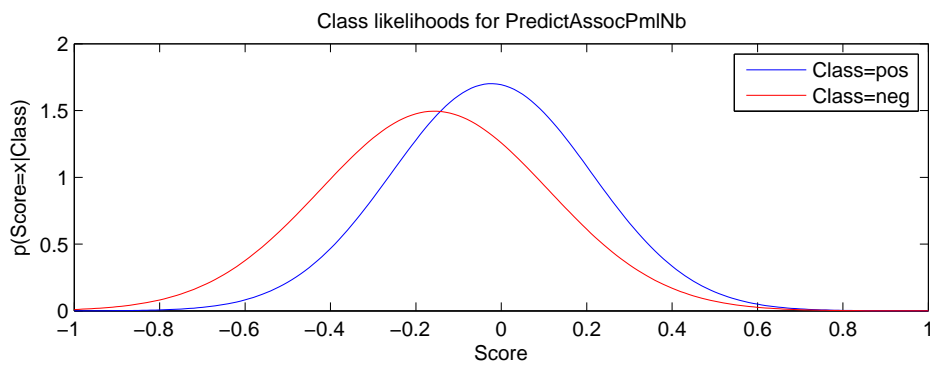


Figure 2: **Gaussian densities estimated from an example SVM trained to predict PML-NB association from sequence.** These class likelihoods can be converted to class posteriors using Bayes' rule, e.g. for the “positive” class ($y = +1$): $p(y = +1|f) = \eta p(f|y = +1)p(y = +1)$, where f is the SVM score (a continuous random variable), and η is a normalization constant.

Table 1: PML-NB associated proteins with extensive protein-protein interactions.

Uniprot	Name	Total	PML-NB	Uniprot	Name	Total	PML-NB
P02340	Tp53	246	18	P26350	Ptma	32	1
P45481	Crebbp	208	13	B1AX90	Tp73	31	7
P48754	Brca1	103	10	Q9QZR5	Hipk2	30	9
P63166	Sumo1	63	13	Q08297	Rad51	30	8
Q60953	Pml	52	17	Q8VEE4	Rpa1	30	6
P23804	Mdm2	50	5	O70133	Dhx9	28	3
Q61026	Ncoa2	46	6	Q9Z265	Chek2	26	6
O35613	Daxx	44	9	Q03347	Runx1	25	1
Q9JIF0	Prmt1	44	5	Q9Z172	Sumo3	19	5
Q9JM05	Pias4	41	2	O88700	Blm	13	6

All proteins with at least 25 interactions. Additionally, Sumo3 and Blm are included (Sumo3 and Blm are hypothesized to be important to the formation of PML-NBs). Uniprot and gene identifiers are specified with number of interactions with the full nuclear proteome ('Total') and number of interactions within the nuclear body ('PML-NB').

Table 2: **Mouse protein data sets.**

<i>Data set name</i>	<i>Size</i>	<i>Definition</i>
Pos-all	76	Mouse nuclear proteins that are orthologs of proteins annotated with the term “PML bodies” in the NPD
Neg-interact	521	Mouse nuclear proteins that are <i>not</i> in Pos-all but that interact with a protein in Pos-all
Neg-1	715	Mouse non-interacting nuclear proteins (1/2)
Neg-2	716	Mouse non-interacting nuclear proteins (1/2)
All-1	1312	Pos-all + Neg-interact + Neg-1
All-2	1313	Pos-all + Neg-interact + Neg-2

Table 3: **The Noisy-OR conditional probability table for the main prediction variable.**

Query protein interacts with					Matches motif		AssocWithPmlNb
Pml	Sumo1	Blm	Rad51	Mdm2	SumoSite	SIM	=T
T	F	F	F	F	F	F	0.38
F	T	F	F	F	F	F	0.25
F	F	T	F	F	F	F	0.67
F	F	F	T	F	F	F	0.47
F	F	F	F	T	F	F	0.20
F	F	F	F	F	T	F	0.09
F	F	F	F	F	F	T	0.07
F	F	F	F	F	F	F	0.00

Parent variables (hubs and sumoylation-related motifs) can be true ('T') or false ('F'). Conditional probabilities are inferred from training data for the combinations shown (excluding the last). The remaining probabilities can be calculated as detailed in Eq. 3. For example, the entry for Pml=T, Sumo=T and the others F, is determined by $1 - (1 - 0.38)(1 - 0.25) = 0.54$. Note that we have left out $P(\text{AssocWithPmlNb} = F | pa(\text{AssocWithPmlNb}))$ since it is simply $1 - P$. The table shown is a representative only and individual models (trained on different data sets) may employ slightly different probabilities.

Table 4: Model and literature evidence for PML-NB predicted proteins.

Prob	Gene	Description and evidence
0.87	Wrn	Werner syndrome ATP-dependent helicase homolog. Model: Interactions with Blm. Has sumoylation site and SIM. Probability of AssocWithPmlNb=true, given only sequence, is 0.34 (this is solely based on the SVM score; see Figure 2 for an illustration). Literature: Endogenous Wrn proteins localize in nucleoli and in PML-NBs (Blander et al., 2002) but the Nuclear Protein Database does not list Wrn as a PML-NB member.
0.83	Mlh1	DNA mismatch repair protein. Model: Interactions with Blm. Has SIM. Probability given sequence is 0.08. Literature: Several DNA damage response proteins reside in PML-NBs (Conlan et al., 2004). Dynamic changes of PML-NBs strongly imply their involvement in DNA repair.
0.79	Smc1	Structural maintenance of chromosomes protein 1A. Model: Interactions with Blm. Has sumoylation site and SIM. Probability given sequence is 0.07.
0.78	Brca2	Breast cancer type 2 susceptibility protein homolog. Model: Interactions with Rad51. Has sumoylation site and SIM. Probability given sequence is 0.70. Literature: Brca2 co-localizes with known PML-NB members Brca1 and Rad51 (Mladenov et al., 2006).
0.78	Prmt8	Protein arginine N-methyltransferase 8. Model: Interactions with none. Has sumoylation site and SIM. Probability given sequence is 1.00. Literature: Its homolog Prmt1 is known to co-localize with PML-NB and regulates DNA repair (Pal and Sif, 2007). Prmt8 is believed to localize to the plasma membrane (Nucprot includes it as nuclear-localized on basis of computational predictions, not experimental evidence).
0.76	Abl1	Proto-oncogene tyrosine-protein kinase. Model: Interactions with Rad51 and Mdm2. Has neither sumoylation site nor SIM. Probability given sequence is 0.06.
0.70	Rb1	Retinoblastoma-associated protein. Model: Interactions with Pml and Mdm2. Has sumoylation site. Probability given sequence is 0.06. Literature: Pml acts with Rb1 and p53 to promote premature senescence. It colocalizes with Pml within nuclear bodies and delocalizes with Pml-Rar-alpha (Alcalay et al., 1998) but is not annotated as a PML-NB protein in the Nuclear Protein Database.
0.70	Sumo2	Small ubiquitin-related modifier 2. Model: Interactions with Pml and Blm. Has sumoylation site. Probability given sequence is 0.52. Literature: Sumo2 is a Sumo isoform and has been shown to co-localize with interphase PML-NBs (Fu et al., 2005).
0.66	Ncor2	Nuclear receptor corepressor 2. Also known as Smrt (silencing mediator for retinoic acid and thyroid hormone receptors). Model: Interactions with Pml. Has sumoylation site and SIM. Probability given sequence is 0.14. Literature: The protein is a matrix-associated deacetylase nuclear body component (Ariumi et al., 2003).
0.65	Tgfbr2	TGF-beta receptor type-2. Model: Interactions with Pml. Has sumoylation site and SIM. Probability given sequence is 0.10. Literature: Cytoplasmic Pml and Daxx have been implicated Tgf-beta signaling (Krieghoff-Henning and Hofmann, 2008).

The table shows all proteins predicted with probability greater than 0.65 to be associated with PML-NBs that are *not* in our positive set. We provide evidence used by the model and independent statements in the literature that are relevant to further corroborate their true localization. Note that several proteins are confirmed to be positives from the literature. Also note that multiple models are collectively queried, leading to slight differences in weighting of features. In particular, the impact of sequence similarities varied between different cross-validation runs.

Table 5: GO terms that are assigned to high-scoring proteins *predicted* to be associated with PML-NBs.

<i>p</i> -value	Term	Proteins	Description
Cellular component			
1.36E-04	GO:0005694*	94	Chromosome
2.26E-05	GO:0016605*	11	PML body
Biological process			
4.62E-06	GO:0044260	281	Cellular macromolecule metabolic process
1.28E-04	GO:0031570*	14	DNA integrity checkpoint
2.77E-04	GO:0006917	34	Induction of apoptosis
1.10E-06	GO:0006464	182	Protein modification process
1.26E-06	GO:0043412	188	Biopolymer modification
2.87E-04	GO:0022402*	123	Cell cycle process
7.02E-05	GO:0006793	122	Phosphorus metabolic process
3.54E-06	GO:0019538	286	Protein metabolic process
1.09E-04	GO:0006281*	73	DNA repair
2.68E-06	GO:0006950*	141	Response to stress
5.23E-06	GO:0006974*	92	Response to DNA damage stimulus
4.62E-06	GO:0044267	281	Cellular protein metabolic process
2.86E-04	GO:0043068	40	Positive regulation of programmed cell death
7.02E-05	GO:0006796	122	Phosphate metabolic process
8.56E-07	GO:0043687	170	Post-translational protein modification
2.78E-04	GO:0008630	14	DNA damage response, signal transduction resulting in induction of apoptosis
1.24E-04	GO:0043065	39	Positive regulation of apoptosis
5.27E-06	GO:0006259*	177	DNA metabolic process
5.97E-06	GO:0009719*	96	Response to endogenous stimulus
2.89E-04	GO:0008629	17	Induction of apoptosis by intracellular signals
9.75E-05	GO:0050896*	210	Response to stimulus
1.08E-05	GO:0006468	99	Protein amino acid phosphorylation
2.21E-04	GO:0016925	6	Protein sumoylation
2.11E-04	GO:0006260*	43	DNA replication
1.26E-05	GO:0016310	101	Phosphorylation
2.77E-04	GO:0012502	34	Induction of programmed cell death
4.30E-05	GO:0042770*	24	DNA damage response, signal transduction
Molecular function			
3.40E-06	GO:0000166	218	Nucleotide binding
1.11E-07	GO:0003824	358	Catalytic activity
9.09E-10	GO:0017076	175	Purine nucleotide binding
2.21E-04	GO:0019789	6	SUMO ligase activity
2.06E-09	GO:0030554	160	Adenyl nucleotide binding
7.59E-06	GO:0043169	318	Cation binding
2.40E-06	GO:0043167	338	Ion binding
2.40E-06	GO:0046872	338	Metal ion binding
2.37E-04	GO:0004871	99	Signal transducer activity
1.80E-05	GO:0046914	293	Transition metal ion binding
2.15E-05	GO:0008270	261	Zinc ion binding
1.63E-05	GO:0016740	154	Transferase activity
1.07E-09	GO:0005524	159	ATP binding
7.96E-05	GO:0016772	120	Transferase activity, transferring phosphorus-containing groups
2.37E-04	GO:0060089	99	Molecular transducer activity
1.08E-04	GO:0016301	107	Kinase activity
3.13E-06	GO:0004672	95	Protein kinase activity
4.74E-06	GO:0005515	629	Protein binding
3.52E-05	GO:0004674	82	Protein serine/threonine kinase activity
5.93E-06	GO:0016773	96	Phosphotransferase activity, alcohol group as acceptor

p-values are uncorrected (after correction $E < 1$). Terms that are over-represented in the set of known PML-NB associated proteins in "All-1" (also $E < 1$) are marked with an asterisk. The third column contains the total number of proteins in "All-1" with that GO term.

Table 6: **The kernel functions with best accuracy for predicting interaction with named hub.**

Hub	Kernel function	AUC
Pml	Local alignment	0.61
Sumo1	Local alignment	0.66
Blm	Mismatch ($k = 3, m = 1$)	0.68
Rad51	Spectrum ($k = 2$)	0.66
Mdm2	Local alignment	0.58

Table 7: Prediction accuracies for representative test scenarios.

Test scenario	AUC
BN has access to interaction and sequence data	0.74
BN has access to interaction data only	0.70
BN has access to sequence data only	0.64
BN has access to sumoylation motif matches only	0.60
BN has access to sequence data and Pml-interactions	0.68
BN has access to sequence data and Sumo1-interactions	0.65
BN has access to sequence data and Blm-interactions	0.64
BN has access to sequence data and Mdm2-interactions	0.64
BN has access to sequence data and Rad51-interactions	0.62
Posterior class probabilities determined from SVM score	0.62
Class probabilities from SVM with logistic function	0.62

All scenarios above the dividing line are generated by the Bayesian network model with different variables instantiated (as indicated). Models below the dividing line represent baseline approaches which rely on the best SVM presented with sequence data only.

Table S1: Strongly predicted nuclear proteins not known to be PML-NB members.

Prob	Uniprot	Name	Description
0.8323	Q9JK91	Mlh1	DNA mismatch repair protein Mlh1
0.7851	Q9CU62	Smc1	Structural maintenance of chromosomes protein 1A
0.7819	P97929	Brca2	Breast cancer type 2 susceptibility protein homolog
0.7756	Q6PAK3	Prmt8	Protein arginine N-methyltransferase 8
0.7629	P00520	Abl1	Proto-oncogene tyrosine-protein kinase ABL1
0.6576	Q9WU42	Ncor2	Nuclear receptor corepressor 2
0.6509	Q62312	Tgfbr2	TGF-beta receptor type-2
0.6466	O08811	Ercc2	TFIIH basal transcription factor complex helicase subunit (DNA repair protein complementing XP-D cells) (DNA excision repair protein ERCC-2)
0.6424	P70270	Rad54l	DNA repair and recombination protein RAD54-like
0.6392	Q61221	Hif1a	Hypoxia-inducible factor 1 alpha
0.6226	P25322	Cnd1	G1/S-specific cyclin-D1
0.6068	P53762	Arnt	Aryl hydrocarbon receptor nuclear translocator (Dioxin receptor, nuclear translocator) (Hypoxia-inducible factor 1 beta)
0.5888	P42227	Stat3	Signal transducer and activator of transcription 3 (Acute-phase response factor)
0.5853	Q5SXJ3	Brip1	Fanconi anemia group J protein homolog (ATP-dependent RNA helicase BRIP1) (BRCA1-interacting protein C-terminal helicase 1) (BRCA1-interacting protein 1) (BRCA1-associated C-terminal helicase 1)
0.5806	Q80X56	Trim69	Tripartite motif-containing protein 69 (RING finger protein 36) (RING finger B-box coiled-coil transcription factor) (Testis-specific RING finger protein)
0.5784	O88907	Pias1	E3 SUMO-protein ligase PIAS1 (Protein inhibitor of activated STAT protein 1) (DEAD/H box-binding protein 1)
0.5753	Q6NZQ4	Paxip1	PAX-interacting protein 1 (PAX transactivation activation domain-interacting protein)
0.5532	O09000	Ncoa3	Nuclear receptor coactivator 3 (Thyroid hormone receptor activator molecule 1) (Receptor-associated coactivator 3) (Amplified in breast cancer-1 protein homolog) (Steroid receptor coactivator protein 3) (CBP-interacting protein)
0.5447	Q64511	Top2b	DNA topoisomerase 2-beta (DNA topoisomerase II, beta isozyme)
0.5424	Q9ERU9	Ranbp2	E3 SUMO-protein ligase (Ran-binding protein 2)
0.5420	P35569	Irs1	Insulin receptor substrate 1
0.5375	Q13547	HDAC1	Histone deacetylase 1
0.5277	Q60698	Ski	Ski oncogene
0.5237	O89090	Sp1	Transcription factor
0.5133	P70365	Ncoa1	Nuclear receptor coactivator 1 (Steroid receptor coactivator 1) (Nuclear receptor coactivator protein 1)
0.5062	Q8CI11	Gnl3	Guanine nucleotide-binding protein-like 3 (Nucleolar GTP-binding protein 3) (Nucleostemin)
0.5007	P11416	Rara	Retinoic acid receptor alpha (Nuclear receptor subfamily 1 group B member 1)

Of a total 53 predictions exceeding a probability of 0.5, 27 are listed. All positives known from the data set (21) and from an initial literature search (5) were removed.