BIOINFORMATICS

Vol. 27 ISMB 2011, pages i7–i14
doi:10.1093/bioinformatics/btr217

Sorting the nuclear proteome

Denis C. Bauer[1,2], Kai Willadsen[2,3], Fabian A. Buske[2], Kim-Anh Lê Cao[4], Timothy L. Bailey[2], Graham Dellaire[5] and Mikael Bodén[2,3,6,*]

[1]Queensland Brain Institute, [2]Institute for Molecular Bioscience, [3]School of Chemistry and Molecular Biosciences, [4]Queensland Facility for Advanced Bioinformatics, The University of Queensland, St Lucia, Australia, [5]Departments of Pathology and Biochemistry & Molecular Biology, Dalhousie University, Halifax, Nova Scotia, Canada and [6]School of Information Technology and Electrical Engineering, The University of Queensland, St Lucia, Australia

ABSTRACT

Motivation: Quantitative experimental analyses of the nuclear interior reveal a morphologically structured yet dynamic mix of membraneless compartments. Major nuclear events depend on the functional integrity and timely assembly of these intra-nuclear compartments. Yet, unknown drivers of protein mobility ensure that they are in the right place at the time when they are needed.

Results: This study investigates determinants of associations between eight intra-nuclear compartments and their proteins in heterogeneous genome-wide data. We develop a model based on a range of candidate determinants, capable of mapping the intra-nuclear organization of proteins. The model integrates protein interactions, protein domains, post-translational modification sites and protein sequence data. The predictions of our model are accurate with a mean AUC (over all compartments) of 0.71.

We present a complete map of the association of 3567 mouse nuclear proteins with intra-nuclear compartments. Each decision is explained in terms of essential interactions and domains, and qualified with a false discovery assessment. Using this resource, we uncover the collective role of transcription factors in each of the compartments. We create diagrams illustrating the outcomes of a Gene Ontology enrichment analysis. Associated with an extensive range of transcription factors, the analysis suggests that PML bodies coordinate regulatory immune responses.

Contact: m.boden@uq.edu.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The cell nucleus has morphologically distinct sub-compartments. However, unlike cytoplasmic organelles, the sub-compartments within the nucleus are not membrane-enclosed, but rather are formed via recruitment of proteins, RNA and DNA.

The translocation of proteins from the cytosol into the nucleus, and their subsequent association with its sub-compartments represent two distinct levels of cellular regulation. At the first level, nuclear import and export of proteins is largely governed by the nuclear-membrane-associated nuclear pore complex (NPC) along with a set of proteins that recognize and actively assist cargo proteins (Stewart, 2007). The import signals used by cargo proteins to bind with carrier proteins were recently extensively explored and modelled (Kosugi *et al.*, 2009). In contrast, at the second level of regulation, the mechanisms underpinning the sorting of proteins into intra-nuclear compartments are not well understood.

The continuous flux of proteins into and out of the nucleus as well as among intra-nuclear compartments underpins central events such as DNA replication, mRNA processing and ribosome biogenesis (Gorski *et al.*, 2006; Misteli, 2007). Hence, to understand these events we need to fully appreciate the mechanisms of intra-nuclear trafficking.

We are now at a stage when we have access to experimental evidence of abundance, localization and modification on a proteomic scale. However, the information offered by many high-throughput techniques does not illustrate the functional purpose of nuclear proteins and compartment structures. Computational modelling, on the other hand, may be able to elucidate functional roles not captured by any individual existing experimental technology (Gorski and Misteli, 2005).

When attempting to gain insight into the underlying mechanisms of translocation, the ability of a model to provide explanations for its predictions is as important as the predictions' accuracy. Several predictors have been reported that evaluate the tendency of a protein to associate with a nuclear compartment (Lei and Dai, 2005; Shen and Chou, 2007). However, these predictors do not provide clear information as to what factors influence these predictions. Additionally, most models are not designed to incorporate any of the constraints that are fundamental to translocation; existing predictors do not recognize that proteins can associate transiently with several compartments at different stages, or that molecular interaction is a prime means of establishing and retaining such associations. Instead, the output of these predictors may be influenced by broad sequence similarity, lacking specific detail of any underlying cause.

This study explores determinants of intra-nuclear compartment association in heterogeneous genome-wide data, and develops a model capable of mapping the intra-nuclear organization of proteins. We use this model to predict the association of the mouse proteome with different intra-nuclear compartments. We also portray transcription factors in terms of how they associate with nuclear architecture to impart novel insights about their 'spatial' synergy. We assign functional roles to individual compartments via their constituent regulatory proteins, using over-represented Gene Ontology (GO) terms in shared target genes.

2 BACKGROUND

We distinguish between eight different locations in the nucleus, each morphologically defined and known to associate with at least 20 different proteins. To select appropriate features for a model,

[*] To whom correspondence should be addressed.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

© The Author(s) 2011. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

we first discuss what is known about the structural and functional role of compartments, their molecular composition and the properties that may modulate protein association.

Association with compartments primarily relies on different post-translational modifications, binding and interaction with core protein members, RNA and regions of DNA, prompting us to peruse a range of different data resources, as discussed in Section 3.

Nucleolus: the largest and best-studied nuclear compartment is the nucleolus. It is highly dynamic and forms at the end of mitosis around a cluster of ribosomal genes, reflecting its primary role of supporting ribosomal biogenesis. Beyond producing ribosomal subunits, nucleoli are involved in cell-cycle control via protein sequestration, and in stress response. Mammalian cells contain one to four nucleoli (see Boisvert *et al.*, 2007 for a review).

Recent large-scale mass spectrometry studies have identified over 700 proteins that stably co-purify with isolated human nucleoli (Andersen *et al.*, 2002, 2005). Several studies have established amino acid motifs that appear to target individual proteins to the nucleolus (Sirri *et al.*, 2008). However, it is unclear whether there is a common sorting principle. Instead it is likely that a multitude of molecular interactions retain proteins within the compartment (Bodén and Teasdale, 2008), which explains the prevalence of WD40 motifs (i.e. scaffolds for protein interactions) in nucleolar proteins (Bickmore and Sutherland, 2002).

Perinucleolar compartment: the perinucleolar compartment (PNC) has been suggested as a pan-cancer marker (Pollock and Huang, 2009) since it appears primarily in transformed and cancer cells, forming a meshwork on the nucleolar surface. PNCs are dynamic structures, assembled in late telophase, and disassembled at the beginning of mitosis (Pollock and Huang, 2009).

Polymerase III transcribed RNA, and proteins involved in RNA metabolism, are known to associate with the PNC. RNA binding domains seem to be critical for the association of some proteins (e.g. Ptb) with the PNC and the compartment integrity is sensitive to the presence of RNase. PNCs appear to be linked to an as yet undefined DNA locus (Pollock and Huang, 2009). Most of the 20 or so proteins that are known to localize to the PNC are also known to associate with other nuclear sites.

Promyelocytic leukaemia body: the Promyelocytic Leukaemia (PML) protein is a core constituent of the nuclear compartment known as the PML nuclear body, or nuclear domain 10 (Bernardi and Pandolfi, 2007). The PML gene was first discovered as the fusion partner of the retinoic acid receptor alpha in a common translocation found in the promyelocytes of patients with acute promyelocytic leukaemia (APL; de Thé *et al.*, 1991). In APL, the PML bodies are absent from leukaemia cells, and PML protein expression is frequently reduced in several forms of cancer, including brain, breast and prostate (Gurrieri *et al.*, 2004). PML bodies are composed primarily of protein, containing little or no DNA or RNA, while they make extensive contacts with chromatin at their periphery (Boisvert *et al.*, 2000). More than 75 proteins have been demonstrated to associate with PML bodies (Dellaire and Bazett-Jones, 2004; Dellaire *et al.*, 2003). Through these protein interactions, PML bodies are thought to function in a host of cellular processes including the anti-viral response, gene regulation, DNA repair, tumour suppression and apoptosis (Bernardi and Pandolfi, 2007). The formation requires the PML protein and is regulated in part by modification of PML by the Small Ubiquitin-like Modifier (SUMO). Many PML body proteins are SUMOylated and/or contain SUMO-interaction motifs (SIMs). In this way PML bodies are thought to form by SUMO-mediated intra- and intermolecular interactions among their components (Shen *et al.*, 2006).

Nuclear speckle: nuclear speckle domains (or Interchromatin Granule Clusters) are believed to be involved in the pre-mRNA processing machinery and regulating factors that are needed for transcription (Lamond and Spector, 2003). As such, these compartments are transit-zones for many RNA binding proteins; a significant portion of speckle proteins exhibit RNA recognition motifs (Bickmore and Sutherland, 2002).

Sometimes referred to as a nuclear speckle targeting signal, many proteins are rich in Arginine and Serine. Indeed, Bickmore and Sutherland (2002) observe that 14 of 18 splicing proteins with an isoelectric point exceeding 10 have an RS domain.

Cajal body: Cajal bodies appear to be sites where proteins associated with a variety of nuclear processes concentrate to increase their functional efficiency, e.g. pre-assembly and modification of spliceosome components (Morris, 2008). Cajal bodies are relatively small and do not occur in all tissue types. Spliceosome formation still occurs, though with lower efficiency, when Cajal bodies are absent, supporting the view that they are non-essential assemblies of cooperating proteins (Morris, 2008).

Coilin is a core Cajal body protein and necessary for recruiting small nuclear ribonucleoproteins, though ‘residual’ bodies still form in its absence (Morris, 2008). Cajal bodies respond dynamically to changes in transcription, and rapid movement of proteins into and out of the body has been observed. They disassemble during mitosis but do not regularly assemble at interphase (Morris, 2008).

Chromatin: chromatin packages DNA and as such is responsible for regulating access of DNA-binding proteins. DNA binding motifs (e.g. so-called AT-hooks) are prevalent in chromatin-associated proteins (Bickmore and Sutherland, 2002). Chromatin consists of many structural proteins, including histones and non-histone proteins, many of which either post-translationally modify histones or remodel chromatin in an ATP-dependent fashion (Becker and Hörz, 2002; Jenuwein and Allis, 2001). Lysines are common in chromatin proteins, and are often modified e.g. acetylated or methylated (Bickmore and Sutherland, 2002). Protein interaction motifs are prevalent in chromatin resident proteins, reflecting the range of interactions required to form and utilize this structure for transcription and replication (Bickmore and Sutherland, 2002).

Nuclear pore complex: the nuclear pore complex (NPC) is a highly structured assembly of approximately 30 proteins that forms a channel through the nuclear membrane (Hetzer *et al.*, 2005). The constituent proteins, nucleoporins, interact in various ways to assist in transporting cargo into and out of the nucleus. Specifically, importins and exportins bind to cargo target sequences via nuclear localization signals and nuclear export sequences, allowing the complex to actively translocate the cargo. This translocation occurs in a multi-stage process involving Ran, which cycles between a GTP- and GDP-bound state (Stewart, 2007).

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

In contrast to other internal compartments, NPCs are relatively static (though some nucleoporins bind and disassociate rapidly), occur in large (though highly variable) numbers, and are found in all cells with a nucleus. They form at the end of mitosis from newly synthesized nucleoporins (Hetzer and Wente, 2009).

NPCs are believed to offer a permissive environment for transcription via DNA, RNA and/or protein interaction, e.g. direct gene interaction or chromatin binding and subsequent modification (Zhao *et al.*, 2009).

Nuclear lamina: in metazoan cells, the nuclear lamina is the inner protein scaffold of the membranous nuclear envelope and is of structural importance (Dechat *et al.*, 2008). The lamina is largely composed of protein filaments (lamins that establish a protein–protein network), is perforated by NPCs, and makes contact with chromatin (Dechat *et al.*, 2008). The lamina is reformed when the nuclear envelope breaks down at cell division (Hetzer *et al.*, 2005).

The lamina is implicated in regulatory processes involving chromatin. Regions of lamina associated with heterochromatin are believed to be transcriptionally repressive, and lamins may play a role in chromatin remodelling and DNA binding. Additionally, abnormalities in the lamina have been linked to specific histone methylations (Dechat *et al.*, 2008).

3 MATERIAL AND METHODS

3.1 Data

We annotated the mouse nuclear proteome with intra-nuclear compartment associations, as described in Mohamad and Bodén (2010). We combined data from multiple data sources including the experimentally determined mouse nuclear proteome (Fink *et al.*, 2008), the Nuclear Protein Database (Dellaire *et al.*, 2003), NOPdb (Leung *et al.*, 2006), and many smaller datasets from the literature. Data from generic (and sometimes automatically annotated) databases such as Uniprot and HPRD was included only when supported by other sources. In addition, we used Ensembl's orthology map to assign annotations for human proteins to the mouse nuclear proteome when necessary.

A total of 3567 proteins are annotated as nuclear, of which 2281 proteins are lacking in any compartment association. The set of 1286 proteins that associate non-exclusively with compartments (see Table 1) are used to train models and establish their test accuracy on held-out subsets (see Section 4.2). We use these models to predict the intra-nuclear compartment association for the set of 2281 proteins lacking this information (see Section 4.3).

Sequence and protein interaction data were sourced from Uniprot and BioGrid (Breitkreutz *et al.*, 2008), respectively. Protein domains were identified using InterProScan and InterPro (Hunter *et al.*, 2009). Sequence motifs and protein post-translational modification sites were identified using the Eukaryotic Linear Motif (ELM) resource (Gould *et al.*, 2010). Combining the information available from these datasets with literature evidence, we determined correlations of protein features grouped into three sets: protein interactions, protein domains and sequence motifs. From this exploration we identify specific features that are used as input features for our model.

3.2 Model

Bayesian networks (BNs) offer a flexible and practical modelling framework, in which nodes represent random variables and directed edges specify their dependencies. Dependencies can be obtained from domain expertise, the incorporation of which results in a graphical representation of the collection of variables, reflecting (potentially causal) relationships between 'parent' and 'child' nodes (where a 'child' variable is conditionally dependent on its 'parent' variables).

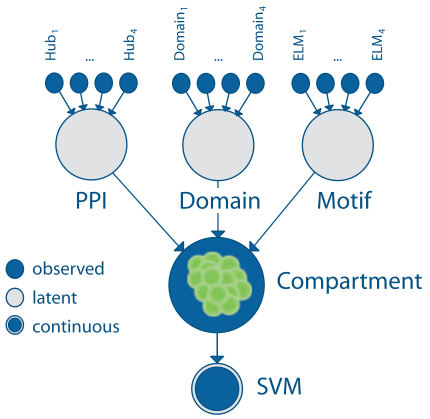

Fig. 1. Template for BN architecture. Nodes are depicted as circles. Edges are directed top-to-bottom, indicating that nodes above are parents to nodes below.

We explored protein features to be used as inputs to a model that is able to identify the compartment(s) to which the protein belongs. In our BN features are random variables. Datasets are thus cast as collections of specific instantiations of random variables, e.g. 'Protein-interacts-with-Pml'=`False`, 'Protein-sequence-has-SUMO-site'=`True` and 'Protein-associates-with-PML-bodies'=`False`.

Each compartment is represented by a set of (non-exclusive) random variables. The nodes for these variables are linked according to a template (see Fig. 1), with one instance of this template for each compartment. The random variables are divided into groups, namely 'PPI', 'Domain' and 'Motif', which are sourced from protein interaction data, InterPro domains and ELM entries, respectively. We expect there to be some dependencies within each group. However, to reduce the number of parameters and increase interpretability, each group of variables is joined via a latent (unobserved) Boolean variable. As a result of training, the latent variable for a group will take a probability that maximizes the likelihood of the data. The latent variables are direct parents of the compartment variable. We thus interpret each group variable as indicating the importance of the values of its features, e.g. if the PPI latent variable is `True`, one or more PPI feature values contribute positively with compartment-specific support. The Boolean compartment variable in the template (in each compartment instance) is a parent of a SVM output variable.

For specific query proteins, variables have known values and can be instantiated. However, the values of some variables are not known; these are left unspecified, and their probabilities are inferred. The joint probability of all variables, $X_1,\ldots,X_N$, in the BN is given by,

$$P(X_1=x_1,\ldots,X_N=x_N)=\prod_{i=1}^{N}P(X_i=x_i\,|\,pa(X_i))$$

where $pa(X_i)$ is the set of parents of the i-th variable.

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. In order to obtain this value, we marginalize over the set of unobserved variables $\mathbf{Y}$,

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

where η is a normalizing constant that ensures that conditional probabilities of X's possible values sum to 1. We use the model to predict the posterior probability of association with a compartment given evidence of the protein (see Section 4.2).

The parameters of the BN are thus the conditional probabilities associated with each node. In this study they are learned from observations in the datasets. The structure of the BN is pre-specified to reflect domain knowledge

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

(see Section 4.1). We use standard Expectation–Maximization to cope with the absence of explicit evidence for variables.

All but one feature in our datasets can be naturally cast as a Boolean random variable. Based on our previous work, to capture similarity between amino acid sequences we train support-vector machines (SVMs) to distinguish between classes of proteins (Bodén *et al.*, 2010). Specifically, for each compartment we train a SVM to distinguish between members and non-members of the compartment. The raw output of the SVM is represented as a continuous random variable, with a Boolean parent variable that is `True` when the protein is a member, and `False` otherwise. This is achieved by parameterizing two Gaussian densities with the means and variances of SVM scores for `True` proteins, and `False` proteins.

4 RESULTS

4.1 Model features

Each compartment is represented by a set of (non-exclusive) random variables incorporated into a Bayesian network, recognizing their dependencies. The nodes for these variables are linked according to a template (see Fig. 1), with one instance of this template for each compartment. The random variables are divided into groups, namely 'PPI', 'Domain' and 'Motif', which are sourced from protein interaction data, InterPro domains and ELM entries, respectively.

By consulting the literature and as discussed above, we identified core members of each compartment as candidates to be used within the PPI group of each compartment module in the BN. For instance, PML bodies use Pml protein as a scaffold for recruiting other members to the compartment (see Section 2). Hence, 'Protein-interacts-with-Pml' is used as a PPI variable in the PML body module. When we failed to establish a set of four candidates, we determined the correlation between their interaction with all other compartment members to nominate additional PPI variables. While not as widely recognized as the manually assigned set, many of the features identified by correlation also have literature support.

Domains from InterPro and motifs from ELM were also identified by consulting the literature. When a domain or motif is known to play an important role for establishing the association of a query protein with the compartment, we use it as a variable. In addition, we observed the correlation between the occurrence of domains in proteins and compartment. We added the domains and motifs with the strongest compartment correlations to the BN, until four InterPro domains and four ELM motifs had been identified for each. The complete set of PPI, domain and motif variables, used in the model experimented within Section 4.2, is provided in Supplementary Table S2 with literature support (when available).

We note that the exploration of candidate features described above may introduce a selection bias. We therefore tried alternative variable selections—randomly choosing what variables to include—and established the overall accuracy of such models. The accuracies changed very little under such perturbation, offering re-assurance that any selection bias is negligible.

4.2 Quantifying predictive accuracy of compartment model

In order to assess the accuracy of our compartment association predictions, we use the area under the ROC curve (AUC) for each compartment classification variable (one versus all). To understand how useful the model is for guiding experimentation, we also provide the AUC50—the AUC measured up until 50 false positives (see Gribskov and Robinson, 1996).

Table 1. Cross-validated prediction accuracy on proteins with known compartment associations

Compartment	Proteins	AUC50 (SD)	AUC (SD)
Cajal body	51	0.22 (0.02)	0.60 (0.03)
Chromatin	323	0.17 (0.02)	0.71 (0.01)
Nuclear lamina	77	0.17 (0.04)	0.70 (0.01)
Nuclear pore	51	0.41 (0.07)	0.79 (0.05)
Nuclear speckle	404	0.24 (0.01)	0.71 (0.01)
Nucleolus	596	0.14 (0.01)	0.60 (0.01)
Perinucleolar	24	0.41 (0.09)	0.80 (0.05)
PML body	91	0.23 (0.06)	0.77 (0.03)
Mean (compartment)		0.25	0.71

Table 1 shows the prediction accuracy (using 5-fold cross-validation averaged over five independent repeats) for a BN that uses protein interactions, domains (from InterPro), post-translational modification sites/sequence motifs (from ELM) and sequence data via SVMs. The BN has a Boolean node for each compartment that is interpreted as the model's prediction (see Section 3). The mean AUC (over all eight compartments) is 0.71. This result is substantially better than random (0.50) and highlights how incomplete annotations are; many compartments have not been subjected to high-throughput screening, and 64% of nuclear proteins have no compartment annotation at all.

We investigated variations to the template module, with and without interactions, domains and motifs (see Supplementary Table S1). We explored the use of gene expression data in the form of DNA microarray data collected over large numbers of tissues (Su *et al.*, 2004) as additional continuous variables to the template. Predictive accuracy varied, and in some cases exceeded that of the standard configuration. Generally, gene expression data contributed very little so we removed it to increase model interpretability. The accuracy using other groups of variables varied by compartment. However, to reduce selection biases we use the standard template configuration consistently in all simulations reported here. This BN performed robustly for all compartments.

4.3 Predicting novel compartment associations

There are many proteins that are known to be imported into the nucleus, but which have no known intra-nuclear compartment association. In this section, we use the model to predict novel compartment associations for the 2281 un-annotated proteins in our assembled dataset.

We create an ensemble model (specified with variables identified in Supplementary Table S2) from the five independently trained models (with different dataset splits). Specifically, we average the posterior probability each model estimates for each compartment, for each protein and also report the standard deviation. We estimate the rate of false discovery (FDR) using the set of proteins with known compartments (samples held out from training).

For each intra-nuclear compartment we report the predictions made by our model up until the FDR exceeds 0.2, given the probability of the prediction (see Supplementary Table S3).

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Table 2. Strongest compartment association prediction of nuclear proteins

Protein	Compartment	Probability	Est. FDR
ZN593 (Q9DB42)	Nucleolus	1.00 (0.00)	0.00
NFAT5 (Q9WV30)	PML body	0.94 (0.00)	0.00
IRF9 (Q61179)	Cajal body	0.94 (0.01)	0.00
PHF12 (Q5SPL2)	Chromatin	0.93 (0.00)	0.00
RUXF (P62307)	Nuclear speckle	0.92 (0.00)	0.12
SMG1 (Q8BKX6)	Nuclear pore	0.90 (0.00)	0.20
ZFHX3 (Q61329)	Nuclear lamina	0.78 (0.01)	0.80
MINT (Q62504)	Perinucleolar	0.40 (0.01)	0.62

In each case, we provide the evidence used by the model to make its prediction; this ability to inspect and interpret predictions is one of the advantages arising from the use of a Bayesian network model. At a confident FDR of 0.2, we predict in addition to those already known, 13 Nucleolar proteins, 13 PML body proteins, 21 Nuclear speckle proteins, 6 Cajal body proteins, 6 Chromatin proteins and 1 Nuclear pore protein.

To first narrow the focus of our discussion, we list the highest-confidence prediction for each compartment in Table 2. For all but one compartment, at least one protein was predicted with a probability higher than 50%; the exception was the Perinucleolar compartment, though due to the small number of perinucleolar proteins in our dataset, the associated prior probability is very small. We also note that the FDR (at the probability of the top prediction) is small for most compartments.

For each of the top predictions, we discuss below details of the evidence used by the model to make the prediction. We use the posterior probability of latent nodes (specific to each group of features in the template module) to indicate the importance of protein interaction, presence of domain and occurrence of motifs for compartment association. In many cases, we also make reference to a specific variable by its name (e.g. a protein name, an InterPro domain name or an ELM name, see Supplementary Table S3 for a complete list with probabilities assigned to each variable). With confident predictions in hand, we conducted a targeted literature search to establish any recent compartment association evidence that has not yet been included in datasets.

Nucleolus: the model predicts with probability 1.0 an association between SSF1 and the nucleolus, confirmed by Kim and Hirsch (1998). With the same probability, the model also predicts ZN593, a zinc finger protein, to be associated with the nucleolus. This prediction is firmly based on protein interactions (the latent variable for the group of protein interactions has a probability of 0.99) with the nucleolar GTP-binding protein 1 (NOG1) and the FHA domain-interacting nucleolar phosphoprotein MKI67.

PML body: the model predicts that NFAT5, a nuclear factor of activated T-cells, is associated with PML bodies due to it containing a p53-like transcription factor (TF) domain (SSF49417) as well as a SUMO-motif, a targeting motif found in a USP7 binding protein, docking to the NTD domain (LIG_USP7_2), and the leucine-rich export signal that binds to CRM1 exportin (TRG_NES_CRM1_1). The latent variable for the domain group of variables has a probability 0.99. This implication of PML bodies in antiviral defence is supported in the literature by (Everett and Chelbi-Alix, 2007). Recently, the related activated T-Cell factor NFAT1 (also known as NFATC2) has been shown to associate with PML NBs, and this interaction appears to enhance the ability of NFAT1 to transactivate its target genes (Lo *et al.*, 2008).

Cajal body: the model predicts IRF9 (0.96), an interferon regulatory factor, to be associated with the Cajal body. The decision was based on it containing the domain of an interferon regulatory factor (PF00605). Furthermore, the decision was greatly influenced by the presence of motifs (0.60), including the major TRAF2-binding consensus motif (LIG_TRAF2_1), the exposed glycosaminoglycan attachment site (MOD_GlcNHglycan) and a subtilisin/kexin isozyme-1 cleavage site (CLV_PCSK_SKI1_1).

Chromatin: both of the top ranking proteins NSD1 and HRX (0.93), are methyltransferases and are confirmed to associate with chromatin (Berdasco *et al.*, 2009; Guenther *et al.*, 2005). The model also predicts PHF12 (0.93), a PHD finger protein, to be associated with chromatin. This is based on interaction with paired amphipathic helix protein Sin3a, on it containing a FYVE/PHD zinc finger domain (SSF57903), the motif recognized by class I SH3 domains (LIG_SH3_1), and on a site for attachment of a fucose residue to serin (MOD_OFUCOSY).

Nuclear speckle: RUXF, a small nuclear ribonucleoprotein, is predicted to be associated with the nuclear speckle based on interaction with survival of motor neuron protein-interacting protein 1 (GEMI2), survival motor neuron protein (SMN) and splicing factor 3A subunit 3 (SF3A3). Additionally, the model draws on the protein containing a GRB2-like Src Homology 2 (SH2) domains binding motif (LIG_SH2_GRB2).

Nuclear pore complex: the Serine/threonine–protein kinase SMG1 is predicted with a probability 0.90 to associate with the nuclear pore. This decision is based on the atypical motif for N-glycosylation site (MOD_N-GLC_2) and the nuclear receptor box motif, which confers binding to nuclear receptors (LIG_NRBOX). Additionally, the high score is partially attributed to the SVM, indicating the presence of currently unknown sequence features.

Nuclear lamina: ZFHX3, a zinc finger homeobox protein, is predicted to associate with the nuclear lamina with a probability 0.78. This is supported by the presence of a motif recognized by class II PDZ domains (LIG_PDZ_2), the atypical motif for N-glycosylation site (MOD_N-GLC_2) as well as a SUMO-motif.

Predicting associations for the full nuclear proteome: only a fraction of all nuclear proteins are so far associated with one or more intra-nuclear location. In response, and to broaden the focus of the discussion, this section estimates the full protein complement of each compartment by identifying the predicted compartment associations (including the possibility of predicting no association) for each of the 2281 unannotated proteins.

In Supplementary Table S4, we publish a comprehensive map of associations with intra-nuclear compartments. We set the probability threshold to be exceeded for a positive prediction for each compartment variable in the Bayesian network using the annotated data. Specifically, to balance sensitivity and specificity, we fix each compartment threshold such that the model renders the correct

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Table 3. Predicted compartment associations for 2281 unannotated nuclear proteins

Compartment	Additional proteins	Probability threshold	FDR at threshold
Cajal body	23	0.24	0.76
Chromatin	509	0.43	0.52
Nuclear lamina	17	0.38	0.68
Nuclear pore	12	0.31	0.43
Nuclear speckle	229	0.41	0.45
Nucleolus	1266	0.44	0.47
Perinucleolar	1	0.29	0.58
PML body	96	0.34	0.64

number of positives on the annotated set. Note that these thresholds render both false negatives and false positives and we use this to estimate the FDR of each model at these permissive thresholds.

Table 3 shows the number of proteins that are predicted as associating with each compartment. The FDR is relatively high at these low probability thresholds, especially for the smaller compartments, which matches the larger number of negatives in the datasets. For reference, Supplementary Table S4 lists all predicted proteins for each compartment.

4.4 Transcription factor analysis

Not unlike how ‘transcription factories’ (Sutherland and Bickmore, 2009) are hypothesized to operate, we seek to establish the collective, regulatory role of individual intra-nuclear compartments. To do so, we first evaluate the prevalence of TFs in each of the compartments. Importantly, we leverage the more comprehensive picture of nuclear organization offered by the model herein and use both known and predicted intra-nuclear associations of nuclear proteins (recall that 2281 of 3567 lack such observations).

We label proteins as TFs by first consulting two sources. RIKEN’s dataset (Kanamori *et al.*, 2004) annotates 1675 mouse proteins to be either TFs or co-regulators of a TF, based on homology to known human TFs and GO annotations. In our data, 977 of these proteins are annotated to be nuclear proteins and 213 are known to associate with at least one compartment. DBD (Wilson *et al.*, 2008) focuses on DNA binding proteins and requires a known DNA-binding/transcriptional regulation domains in the protein sequence. DBD contains 2549 predicted mouse TFs, of which 775 are annotated to be nuclear and only 95 have a known compartment. Around 542 nuclear proteins are annotated by both datasets to be TFs, and 66 of these have a known compartment.

Table 4 lists the numbers of TFs among proteins in the different compartments when using RIKEN’s and DBD’s TF definitions. About 17% and 7% of the proteins with known compartment association are identified as TFs (for RIKEN and DBD, respectively). Using this as a background, we can measure TF statistical enrichment for each compartment by applying the Fisher Exact test to the counts of TFs and non-TFs in a compartment and not in the compartment. According to this analysis, chromatin and PML body are significantly enriched in TFs (discussed below), whereas nuclear lamina and nucleolus are under-enriched ($P \ll 0.05$). Reassuringly, no TF was found to be associated with the nuclear pore. The same trend is observed with both the RIKEN and DBD definitions. It is worth noting that nuclear speckles appear to be close to average and are thus not significantly enriched in TFs—contrary to expectations (see Section 2).

Table 4. Transcription factor enrichment in compartments, with significant over-representation within compartments marked

Compartment	RIKEN TFs		DBD TFs	
	Count	Percentage	Count	Percentage
Cajal body	10	19.6	4	7.8
Chromatin	100	31.0*	47	14.6*
Nuclear lamina	4	5.2	4	5.2
Nuclear pore	0	0.0	0	0.0
Nuclear speckle	54	13.4	19	4.7
Nucleolus	71	11.9	18	3.0
Perinucleolar	6	25.0	0	0.0
PML body	27	29.7*	17	18.7*
All	213	16.9	95	7.4

*$P < 0.05$.

Previous work indicates a potential gene regulatory role of PML bodies (Block *et al.*, 2006; Lallemand-Breitenbach and de Thé, 2010), beyond that of acting as a site for post-translational modifications (Gupta *et al.*, 2008; Song *et al.*, 2008). For instance, Wang *et al.* (2004) observe that PML bodies localize to sites of high transcriptional activity. Also, PML bodies are known to sequester TFs for timely release (Lin *et al.*, 2003). Our analysis lends support to such theories by identifying a large set of TFs that we predict or experimentally observe to associate with PML bodies.

Intra-nuclear co-localization of transcription factors may indicate that the factors co-operate, or that the localization site itself has a regulatory role. To investigate if a nuclear compartment contributes to transcriptional regulation, we identified the set of TFs with a DNA binding site in TRANSFAC (Matys *et al.*, 2006) that are known or predicted to associate with each compartment. By scoring the presence of binding sites in all promoters in mouse, we determined the statistical enrichment of GO terms for the putative targets of the applicable group of TFs; note that the set of enriched GO terms are those associated with the putative targets, not those associated with the TFs themselves. Enrichment is established through the same statistically rigorous method as developed in our previous work (Buske *et al.*, 2010). Importantly, this test does not require that sites are co-bound, but instead only considers whether the set of potential gene products are annotated in ways that cannot be explained by chance alone.

For each compartment *c*, in Supplementary Figures S1–S7, we present all GO terms that are statistically over-represented in gene targets of member TFs (see Supplementary Table S5 for a list of TF binding matrices used to establish putative targets for each compartment). Diagrams contain terms taken from the Biological Process and Molecular Function ontologies. We note that many essential and specific terms are only supported when TFs that are predicted to belong to the compartment are included. With PML bodies particularly enriched in TFs, to illustrate the utility of our method, we discuss in more detail the GO terms identified for this compartment below.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

The TFs that localize to PML bodies target genes that are associated (in a statistically significant sense) with a variety of biological processes. In particular, we note that TFs in PML bodies target immune genes. Several terms relate broadly to positive regulation of immune response, cytokine production, leukocyte and lymphocyte mediated immunity. These are observations identified via predicted PML body members, that are well supported in the literature. For instance, the morphology of PML bodies changes markedly in response to interferon, a protein with viral defence and immunomodulatory activities (Regad and Chelbi-Alix, 2001). In fact, interferon directly induces transcription of core PML body proteins (Everett and Chelbi-Alix, 2007). We also note that targets involved in the regulation of apoptosis/cell death and cytotoxicity figure strongly in this analysis of PML body TFs. The role of PML bodies in apoptosis is well-established (Bernardi *et al.*, 2008). The analysis suggests that PML bodies play a signal transduction role involving G-protein coupled receptors, e.g. cytokines. There are some indications in the literature that several PML body proteins are targets of cytokines (Salomoni, 2009).

5 CONCLUSION

We introduce a computational model that integrates evidence garnered from protein interaction, domain and post-translational modification data, to systematically address fundamental questions of nuclear compartment localization. We use a Bayesian Network, a probabilistic and transparent modelling framework, whose decisions can clearly be traced back to the influence of the provided biological features.

After carefully designing and training our model, we identify and make available the determinants of intra-nuclear compartment association of the full mouse nuclear proteome. The model predicts intra-nuclear compartment associations of each protein with an accuracy substantially above chance (AUC is 0.71) and indicates the individual factors that influence its decisions. The mobility of a nuclear protein is typically determined by its interaction with other nuclear components and is often modulated by post-translational modifications, including sumoylation. We also note that specific functional domains are sometimes enriched in compartments and can be used to determine likely associations.

This work presents a unique intra-nuclear protein association map involving all the main compartments for 3567 mouse proteins, and provides detailed justifications for each individual prediction. This resource will enable cell biologists to effectively investigate what compartments a protein of interest is likely to associate with, and identify the factors including interactions and domain features that modulate this sorting or at least unify proteins from the same compartment.

Whether nuclear compartments in general fill a regulatory role remains an open question—but our analyses lend significant support to the idea that PML bodies and chromatin are implicated in regulatory processes. We offer an analysis that is not simply based on individual TFs, but one that groups them according to compartment association to identify their downstream targets. We publish detailed Gene Ontology diagrams for each compartment that will allow biologists to investigate the statistical support for any regulatory function and to guide targeted experimentation.

6 SUPPLEMENTARY MATERIAL

Supplementary Table S1 shows the prediction accuracy for different BN architectures. Supplementary Table S2 lists the PPI, domain and motif variables provided to the model from biological evidence. Supplementary Table S3 contains the top-25 predictions for each compartment, the basis for the predictions and supporting evidence from the literature where available. Supplementary Table S4 lists compartment predictions for all proteins. Supplementary Table S5 provides the list of compartment-associated TFs with their TRANSFAC binding motifs.

Supplementary Figures S1–S7 show trees of GO terms that are over-represented in the set of target genes of TFs associated with a given intra-nuclear compartment. Terms are coloured green if they can be considered over-represented using only the known set of compartment-associated TFs. Terms coloured blue are only discovered when using both known and predicted TFs; thus, blue terms constitute novel functional predictions made on the basis of results from our model. Uncoloured terms have been included only to provide context based on the GO term hierarchy.

ACKNOWLEDGEMENT

We thank Nurul Mohamad for collecting the datasets that enabled the development of the model.

Funding: This study received support from the Australian Research Council (ARC) Centre of Excellence in Bioinformatics. G.D. is a Canadian Institutes of Health Research (CIHR) New Investigator and is funded by an operating grant from the CIHR (MOP-84260).

Conflict of Interest: none declared.

REFERENCES

Andersen,J.S. *et al*. (2002) Directed proteomic analysis of the human nucleolus. *Curr. Biol.*, **12**, 1–11.

Andersen,J.S. *et al*. (2005) Nucleolar proteome dynamics. *Nature*, **433**, 77–83.

Becker,P.B. and Hörz,W. (2002) ATP dependent nucleosome remodeling. *Annu. Rev. Biochem.*, **71**, 247–273.

Berdasco,M. *et al*. (2009) Epigenetic inactivation of the sotos overgrowth syndrome gene histone methyltransferase NSD1 in human neuroblastoma and glioma. *Proc. Natl Acad. Sci. USA*, **106**, 21830–21835.

Bernardi,R. and Pandolfi,P.P. (2007) Structure, dynamics and functions of promyelocytic leukaemia nuclear bodies. *Nat. Rev. Mol. Cell Biol.*, **8**, 1006–1016.

Bernardi,R. *et al*. (2008) Regulation of apoptosis by PML and the PML-NBs. *Oncogene*, **27**, 6299–6312.

Bickmore,W. and Sutherland,H. (2002) Addressing protein localization within the nucleus. *EMBO J.*, **21**, 1248–1254.

Block,G.J. *et al*. (2006) Transcriptional regulation is affected by subnuclear targeting of reporter plasmids to PML nuclear bodies. *Mol. Cell Biol.*, **26**, 8814–8825.

Bodén,M. and Teasdale,R.D. (2008) Determining nucleolar association from sequence by leveraging protein-protein interactions. *J. Comput. Biol.*, **15**, 291–304.

Bodén,M. *et al*. (2010) A Bayesian network model of proteins' association with Promyelocytic leukemia (PML) nuclear bodies. *J. Comput. Biol.*, **17**, 617–630.

Boisvert,F.M. *et al*. (2000) Promyelocytic leukemia (PML) nuclear bodies are protein structures that do not accumulate RNA. *J. Cell Biol.*, **148**, 283–292.

Boisvert,F.-M. *et al*. (2007) The multifunctional nucleolus. *Nat. Rev. Mol. Cell Biol.*, **8**, 574–585.

Breitkreutz,B.-J. *et al*. (2008) The BioGRID interaction database: 2008 update. *Nucleic Acids Res.*, **36**, D637–D640.

Buske,F.A. *et al*. (2010) Assigning roles to DNA regulatory motifs using comparative genomics. *Bioinformatics*, **26**, 860–866.

Dechat,T. *et al*. (2008) Nuclear lamins: major factors in the structural organization and function of the nucleus and chromatin. *Genes Dev.*, **22**, 832–853.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Dellaire,G. and Bazett-Jones,D.P. (2004) PML nuclear bodies: dynamic sensors of DNA damage and cellular stress. *Bioessays*, **26**, 963–977.

Dellaire,G. *et al*. (2003) The nuclear protein database (NPD): sub-nuclear localization and functional annotation of the nuclear proteome. *Nucleic Acids Res.*, **31**, 328–330.

de Thé,H. *et al*. (1991) The PML-RARα fusion mRNA generated by the t(15;17) translocation in acute promyelocytic leukemia encodes a functionally altered RAR. *Cell*, **66**, 675–684.

Everett,R.D. and Chelbi-Alix,M.K. (2007) PML and PML nuclear bodies: implications in antiviral defence. *Biochimie*, **89**, 819–830.

Fink,J.L. *et al*. (2008) Towards defining the nuclear proteome. *Genome Biol.*, **9**, R15.

Gorski,S.A. *et al*. (2006) The road much traveled: trafficking in the cell nucleus. *Curr. Opin. Cell Biol.*, **18**, 284–290.

Gorski,S. and Misteli,T. (2005) Systems biology in the cell nucleus. *J. Cell Sci.*, **118**(Pt 18), 4083–4092.

Gould,C.M. *et al*. (2010) ELM: the status of the 2010 eukaryotic linear motif resource. *Nucleic Acids Res.*, **38**, D167–D180.

Gribskov,M. and Robinson,N.L. (1996) Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comput. Chem.*, **20**, 25–33.

Guenther,M.G. *et al*. (2005) Global and hox-specific roles for the MLL1 methyltransferase. *Proc. Natl Acad. Sci. USA*, **102**, 8603–8608.

Gupta,P. *et al*. (2008) Retinoic acid-stimulated sequential phosphorylation, PML recruitment, and sumoylation of nuclear receptor TR2 to suppress Oct4 expression. *Proc. Natl Acad. Sci. USA*, **105**, 11424–11429.

Gurrieri,C. *et al*. (2004) Loss of the tumor suppressor PML in human cancers of multiple histologic origins. *J. Natl Cancer Inst.*, **96**, 269–279.

Hetzer,M.W. and Wente,S.R. (2009) Border control at the nucleus: biogenesis and organization of the nuclear membrane and pore complexes. *Dev. Cell*, **17**, 606–616.

Hetzer,M.W. *et al*. (2005) Pushing the envelope: structure, function, and dynamics of the nuclear periphery. *Annu. Rev. Cell Dev. Biol.*, **21**, 347–380.

Hunter,S. *et al*. (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211–D215.

Jenuwein,T. and Allis,C.D. (2001) Translating the histone code. *Science*, **293**, 1074–1080.

Kanamori,M. *et al*. (2004) A genome-wide and nonredundant mouse transcription factor database. *Biochem. Biophys. Res. Commun.*, **322**, 787–793.

Kim,J. and Hirsch,J.P. (1998) A nucleolar protein that affects mating efficiency in Saccharomyces cerevisiae by altering the morphological response to pheromone. *Genetics*, **149**, 795–805.

Kosugi,S. *et al*. (2009) Six classes of nuclear localization signals specific to different binding grooves of importin alpha. *J. Biol. Chem.*, **284**, 478–485.

Lallemand-Breitenbach,V. and de Thé,H. (2010) PML nuclear bodies. *Cold Spring Harb. Perspect Biol.*, **2**, a000661.

Lamond,A.I. and Spector,D.L. (2003) Nuclear speckles: a model for nuclear organelles. *Nat. Rev. Mol. Cell Biol.*, **4**, 605–612.

Lei,Z. and Dai,Y. (2005) An SVM-based system for predicting protein subnuclear localizations. *BMC Bioinformatics*, **6**, 291.

Leung,A.K.L. *et al*. (2006) NOPdb: Nucleolar Proteome Database. *Nucleic Acids Res.*, **34** (Suppl. S1), D218–D220.

Lin,D.-Y. *et al*. (2003) Promyelocytic leukemia protein (PML) functions as a glucocorticoid receptor co-activator by sequestering Daxx to the PML oncogenic domains (PODs) to enhance its transactivation potential. *J. Biol. Chem.*, **278**, 15958–15965.

Lo,Y.-H. *et al*. (2008) Selective activation of nfat by promyelocytic leukemia protein. *Oncogene*, **27**, 3821–3830.

Matys,V. *et al*. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.

Misteli,T. (2007) Beyond the sequence: cellular organization of genome function. *Cell*, **128**, 787–800.

Mohamad,N. and Bodén,M. (2010) The proteins of intra-nuclear bodies: a data-driven analysis of sequence, interaction and expression. *BMC Syst. Biol.*, **4**, 44.

Morris,G.E. (2008) The Cajal body. *Biochim. Biophys. Acta*, **1783**, 2108–2115.

Pollock,C. and Huang,S. (2009) The perinucleolar compartment. *J. Cell Biochem.*, **107**, 189–193.

Regad,T. and Chelbi-Alix,M.K. (2001) Role and fate of PML nuclear bodies in response to interferon and viral infections. *Oncogene*, **20**, 7274–7286.

Salomoni,P. (2009) Stemming out of a new PML era? *Cell Death Differ.*, **16**, 1083–1092.

Shen,H.-B. and Chou,K.-C. (2007) Nuc-ploc: a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. *Protein Eng. Des. Sel.*, **20**, 561–567.

Shen,T.H. *et al*. (2006) The mechanisms of PML-nuclear body formation. *Mol. Cell*, **24**, 331–339.

Sirri,V. *et al*. (2008) Nucleolus: the fascinating nuclear body. *Histochem. Cell Biol.*, **129**, 13–31.

Song,M.S. *et al*. (2008) The deubiquitinylation and localization of PTEN are regulated by a HAUSP-PML network. *Nature*, **455**, 813–817.

Stewart,M. (2007) Molecular mechanism of the nuclear protein import cycle. *Nat. Rev. Mol. Cell Biol.*, **8**, 195–208.

Sutherland,H. and Bickmore,W.A. (2009) Transcription factories: gene expression in unions? *Nat. Rev. Genet.*, **10**, 457–466.

Su,A.I. *et al*. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.

Wang,J. *et al*. (2004) Promyelocytic leukemia nuclear bodies associate with transcriptionally active genomic regions. *J. Cell Biol.*, **164**, 515–526.

Wilson,D. *et al*. (2008) DBD–taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res.*, **36**, D88–D92.

Zhao,R. *et al*. (2009) Nuclear neighborhoods and gene expression. *Curr. Opin. Genet. Dev.*, **19**, 172–179.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016