

Determining nucleolar association from sequence by leveraging protein-protein interactions

Mikael Bodén^{a,b} and Rohan D. Teasdale^a

^aARC Centre of Excellence in Bioinformatics and
Institute for Molecular Bioscience,

^bSchool of Information Technology and Electrical Engineering,
University of Queensland, QLD 4072, Australia.

m.boden@uq.edu.au

October 17, 2007

Abstract

Controlled intra-nuclear organisation of proteins is critical for sustaining correct function of the cell. Proteins and RNA are transported by passive diffusion and associate with compartments by virtue of diverse molecular interactions—presenting a challenging problem for data-driven model-building. An increasing inventory of proteins with known intra-nuclear destination and proliferation of molecular interaction data motivate an integrative method, leveraging the existing evidence to build accurate models of intra-nuclear trafficking.

Kernel Canonical Correlation Analysis (KCCA) enables the construction of predictors which are based on genomic sequence data but leverages other knowledge sources during training. The approach specifically involves the induction of protein

sequence features and relations most pertinent to the recovery of nucleolar associated protein-protein interactions. With success rates of about 78%, the classification of nucleolar association from KCCA-induced features surpasses that of baseline approaches. We observe that the coalescence of protein-protein interaction data with sequence data enhances the prediction of highly interconnected, key ribosomal and RNA-related nucleolar proteins.

Supplementary material: <http://www.itee.uq.edu.au/~pprowler/nucleoli>.

1 Introduction

Intra-nuclear protein trafficking plays a profound role sustaining the function and integrity of a eukaryotic cell, e.g. biogenesis and gene regulation rely on the presence of specific proteins (Carmo-Fonseca, 2002; Stein *et al.*, 2000). In contrast to cytoplasmic organelles, intra-nuclear compartments are not membrane-bound and form protein destinations only by virtue of molecular associations. Microscopic techniques have revealed that passive diffusion and binding interactions best describe how proteins (and RNA) move about in the nucleus and how such intra-nuclear compartments preserve their state (Misteli, 2001). This paper evaluates a method to model interaction-based associations with protein destinations representing a novel approach for predicting localisation.

The nucleolus is a large compartment inside the nucleus of a eukaryotic cell. Several significant functions occur in the nucleolus, including ribosomal RNA synthesis, processing, and assembly into subunits, but also pre-assembly of the signal recognition particle, DNA repair and tRNA maturation. Additionally, several cell-cycle regulated proteins seem to be kept in (and released by) the nucleolus (Leung *et al.*, 2003). The nucleolus is thus a heterogeneous molecular aggregate involving a variety of components, e.g. rRNA genes, rRNAs, ribosome-related house-keeping proteins and enzymes, and ribosomal protein subunits. Almost of quarter of proteins found

in the nucleolus bind to DNA (Leung *et al.*, 2003). Many nucleolus-associated proteins also associate with other intra-nuclear compartments, e.g. the Cajal bodies and Paraspeckles. Some proteins are stably associated with the nucleolus and others enter and exit the nucleolus with diverse kinetics. Moreover, the protein content varies with metabolic and cell-cycle conditions (Andersen *et al.*, 2005; Leung *et al.*, 2003; Dundr and Misteli, 2001; Misteli, 2001).

Recent experimental efforts using mass-spectrometry have identified almost 700 proteins from the human nucleolar proteome of which about 30% are uncharacterised (Andersen *et al.*, 2005). There is now a great need to leverage this near-complete inventory to characterise the function of the nucleolus (Hinsby *et al.*, 2006) and to model the fluid participation of components of the nucleolus and other intra-nuclear compartments.

In this paper we introduce protein-protein interaction data as an additional resource which a sequence-based predictive model can exploit. The interactions represent the means by which proteins associate with the intra-nuclear molecular aggregate that make up the compartment. Specifically, we ask if we can predict association to the nucleolus by virtue of interactions with known nucleolar proteins. The approach involves the induction of protein sequence features and relations most pertinent to the recovery of nucleolar associated protein-protein interactions and then using these features to predict protein nucleolar association. We employ and evaluate a kernel-based framework adapted from Vert and Kanehisa (2003) and Yamanishi *et al.* (2004) to enable integration of heterogeneous genomic data (sequence and protein-protein interaction network data) into a single predictive model.

The prediction of association with intra-nuclear compartments has not been explored extensively in the literature. Bickmore and colleagues (Bickmore and Sutherland, 2002; Dellaire *et al.*, 2003) illustrate how a number of physico-chemical protein properties weakly correlate with intra-nuclear organisation and present the Nuclear Protein Database (NPD) with nuclear proteins with known compartment (<http://npd.hgu.mrc.ac.uk/>).

Using the NPD, Lei and Dai (2005, 2006) developed a predictor using machine learning of six different nuclear compartments including the nucleolus. Multi-compartmental proteins were removed from the data set (prior to training) to avoid the ambiguous presentation of data while training their classifiers. In their most refined model, there is a Gene Ontology (GO) module which relies on the identification of GO terms of the protein and its homologs. A separate support-vector machine is trained to map the sequence to one of the six classes. The GO term module elevates overall performance considerably (the correlation coefficient for nucleolus improves from 0.37 to 0.66). However, collected GO terms (a) sometimes include terms explicitly identifying localisation and (b) always need to be known in advance, compromising the predictor’s utility.

Hinsby *et al.* (2006) devised a system from which novel nucleolar proteins could be semi-automatically identified. By cross-checking protein-protein interactions involving known nuclear proteins with mass spectrometry data of the nucleolus, they ranked nucleolar protein complexes. Then, by targeted search for 55 candidates in the raw mass spectrometry data, eleven novel nucleolar proteins were fingered. Their study indicates the potential of the approach taken herein: assigning intra-nuclear compartment membership in terms of interactions with residents rather than possibly elusive compartment-unifying features.

2 Method

Since kernel methods assume access to samples only through their pairwise inner products, each such computation can be replaced by a call to a kernel function, $k(\cdot, \cdot)$. Formally, a valid kernel on an arbitrary space X can always be represented as an inner product in a reproducing kernel Hilbert space F (RKHS; also known as the kernel’s feature space) (Schölkopf and Smola, 2002). That is, when samples $x_1, x_2, \dots, x_N \in X$, then $k(x_i, x_j) = \phi(x_i)^T \phi(x_j)$, for any i and j , and $\phi(\cdot)$ is the

projection of samples to F . As exploited below, the kernel function can sometime avoid computing the inner product explicitly and instead rely on domain-specific methods for evaluating a similarity between pairs of samples (Schölkopf *et al.*, 1998).

In this study we use two kernel methods: the support-vector machine (Vapnik, 1998) and Kernel Canonical Correlation Analysis (KCCA) (Akaho, 2001; Bach and Jordan, 2002). Originally due to Hotelling, Canonical Correlation Analysis (CCA) finds a linear transformation of pairs of numeric vectors such that their correlation is maximised. KCCA extends CCA to operate on $x'_A \in X_A$ and $x'_B \in X_B$ when there is a kernel $k_A : X_A \times X_A \mapsto \mathfrak{R}$ and a kernel $k_B : X_B \times X_B \mapsto \mathfrak{R}$. We will refer to their intermediate kernel feature spaces as F_A and F_B , respectively. The use of kernels enables the exploration of non-linear relations between the two sample spaces and the use of non-numeric data (like sequences and interaction networks).

Implicitly, KCCA finds $\mathbf{w}_A \in F_A$ and $\mathbf{w}_B \in F_B$ such that Equation 1 is maximised over all sample pairs $(x'_A, x'_B) \in (x_{A,1}, x_{B,1}), \dots, (x_{A,N}, x_{B,N})$.

$$\frac{\text{Cov}(\langle \mathbf{w}_A, \phi_A(x'_A) \rangle, \langle \mathbf{w}_B, \phi_B(x'_B) \rangle)}{\sqrt{\text{Var}(\langle \mathbf{w}_A, \phi_A(x'_A) \rangle) \text{Var}(\langle \mathbf{w}_B, \phi_B(x'_B) \rangle)}} \quad (1)$$

Multiple features in descending order of correlation are usually sought this way. We refer to the vectors formed by concatenating L feature-values as $\mathbf{f}_A(x'_A)$ and $\mathbf{f}_B(x'_B)$, for any $x'_A \in X_A$ and $x'_B \in X_B$, respectively. Regularisation and orthogonality conditions apply to avoid feature-vectors that overfit the training data and to promote components with high information content (Akaho, 2001; Bach and Jordan, 2002). In practise, the solution of \mathbf{f}_A and \mathbf{f}_B is given by the functions corresponding to the largest eigenvalues of a generalised eigenvalue problem (Akaho, 2001; Bach and Jordan, 2002).

KCCA extracts features of the involved samples spaces that enable their coalescence. The cross-space similarity between a sample $x'_A \in X_A$ and any sample $x'_B \in X_B$ can be quantified by simply determining the correlation between

$\mathbf{f}_A(x'_A)$ and $\mathbf{f}_B(x'_B)$, where higher correlation means more similar. The weight vector \mathbf{w}_A is not available explicitly, so we use the method of Yamanishi *et al.* (2004, p. 1367) to determine the l th correlated value for a novel sample as $f_A^l(x'_A) = \sum_{i=1}^N \alpha_A^l(x_{A,i}) \cdot k(x_{A,i}, x'_A)$, where $\alpha_A^l(x_{A,i})$ is the l th element in the eigenvector for $x_{A,i}$ produced as a result of the aforementioned eigenvalue problem. Note that in all simulations reported on herein the kernels are standardly normalised such that $k(x, x) = 1$ for all x (Schölkopf and Smola, 2002).

Yamanishi *et al.* (2004) successfully inferred parts of a protein-protein interaction network by employing KCCA. They used the so-called diffusion kernel (Kondor and Lafferty, 2002) on a smaller “gold standard” network of yeast proteins combined with kernels for several heterogeneous but noisy data sources (including gene expression data). Using cross-validation on the gold standard network, it was noted that the addition of several data sources increased the overall ability to infer protein-protein interactions. A range of alternative unsupervised network inference strategies were shown to be inferior. Recently, Kato *et al.* (2005) developed and evaluated a variation of KCCA based on expectation-minimization for inferring missing entries in a network. Geurts *et al.* (2007) also proposed an alternative approach for network inference. Their method is based on output kernel trees. The performances of the alternative methods were both on par with KCCA.

We choose kernel functions to reflect the nature of the problem domain as outlined below: a network kernel for embedding protein-protein interactions and a sequence kernel for embedding protein sequences in the two kernel feature spaces to be processed by KCCA.

2.1 Protein-protein interaction kernel

The recovery (or discovery) of protein-protein interaction pairs involves the properties between, rather than about, individual samples. Model-driven kernels contain

knowledge about the input space, i.e. about relationships between samples. We use the diffusion kernel that defines a feature space for a network of samples (Kondor and Lafferty, 2002).

The diffusion kernel probes the edges connecting all samples. The kernel matrix (all pairwise kernel evaluations) is defined as $K_d = \exp(\beta(A - D))$ where A is the adjacency matrix (a binary matrix indicating which pairs of samples are directly connected) and D is the diagonal matrix of sample connectivity (each sample's absolute number of connections). A positive β represents the diffusion bandwidth (0 means none).

To illustrate the view of the network provided by the diffusion kernel, we arbitrarily identified three protein complexes involved in the nucleolus (taken from Hinsby *et al.* (2006)). The protein-protein interaction network for the complexes is shown together with the resulting kernel matrix in Figure 1. Kernel evaluations between samples in disjoint networks (like the complexes or single non-interacting proteins) are always 0. Kernel evaluations between samples in the same network are positive to a degree determined by their relative closeness (and the globally defined diffusion bandwidth).

---Figure 1 about here---

2.2 Amino acid sequence kernel

Syntax-driven kernels allow the tailoring of feature spaces for domain-specific data structures, e.g. sequences and trees. For sequence data, we explore spectrum-based kernels (Leslie and Kuang, 2004) which capitalise on shared, short sub-sequences and the Local Alignment kernel (Saigo *et al.*, 2004) which essentially scores the extent with which two sequences align. Specifically, we use the Spectrum kernel, the Mismatch kernel, the Wildcard kernel, and the Local Alignment kernel.

We illustrate the Wildcard and Local Alignment kernel views of the data for pro-

teins participating in the three nucleolar complexes in Figure 2. Visual inspection of Figure 2 (relative Figure 1) reveals that some intra-complex similarities exist but that inter-complex discrimination is non-obvious. Note that the point is not necessarily to discriminate between individual complexes. Rather, KCCA is expected to find sequence features that coalesce with protein-protein interactions and vice versa.

---Figure 2 about here---

2.3 Models

After training, KCCA provides us with two sets of feature vectors, one set for sequences and one set for network vertices. Each sample pair (x'_A, x'_B) thus corresponds to two feature vectors, $\mathbf{f}_A(x'_A)$ and $\mathbf{f}_B(x'_B)$, ideally correlated with one another.

2.3.1 Inferring protein-protein interaction

Test samples do not have interaction data attached, thus we find ways of inferring such information from the sequence projection, \mathbf{f}_A . Assuming successful training, for a query protein x'_A , we seek out the feature vector i in the alternate space (defined by \mathbf{f}_B) with the highest correlation (Yamanishi *et al.*, 2004).

$$\arg \max_i \frac{Cov(\mathbf{f}_A(x'_A), \mathbf{f}_B(x_{B,i}))}{\sqrt{Var(\mathbf{f}_A(x'_A))Var(\mathbf{f}_B(x_{B,i}))}} \quad (2)$$

Analogously, we determine the n most correlated samples in the alternate sample space. Subsequently, by look-up, the set of interactions exhibited by the correlated samples is viewed collectively as the predicted interaction set for the novel sample. For example, if protein A interacts with B, and a novel protein C is correlated with A, the prediction is that C interacts with B since C evidently shares interaction features with the “proxy” protein A.

To gauge the prediction strength of interaction, each interaction partner is associated with the specific correlation that links it to the query protein. If the same interaction partner is identified by several proxies, the prediction strength is the sum of correlation coefficients.

In KCCA, λ_A and λ_B controls regularisation of \mathbf{f}_A and \mathbf{f}_B , respectively. We initially grid searched their settings to determine configurations with good performance. Parameter sweeps are presented in the supplementary material. $\lambda_A = 0.9$ and $\lambda_B = 1.3$ rendered good and stable network inference performance.

2.3.2 Classifying nucleolar association

We use three methods for classifying a protein to be nucleolus-associated. The first method initially identifies the n most correlated proteins per Equation 2. The class membership of the query protein is then based on the class membership (+1 or -1) of each its correlates, weighted by the corresponding correlation coefficient (the test is positive if and only if the sum is positive). After some preliminary trials, we decided to use $n = 10$ in all tests. $\lambda_A = 0.7$ and $\lambda_B = 0.1$ rendered good and stable classification performance. Sweeps are presented in the supplementary material.

Inspired by the models of Lei and Dai (2005), the second method serves as a baseline for comparison as it excludes KCCA entirely. It uses only a support-vector machine which takes as input any sequence data, x'_A and uses a sequence kernel. We trial all aforementioned sequence kernels.

The third method is based on a support-vector machine equipped with a simple linear kernel which takes as input the vector $\mathbf{f}_A(x'_A)$, where x'_A is the query protein. The process follows the recipe used by Vert and Kanehisa to classify DNA microarray data in the context of biological pathway data encoded by a diffusion kernel. We also present the query protein through a sequence kernel and simply combine the two kernel evaluations by their product (which is a valid composite kernel). Again, we initially searched λ -values, finding λ_A and λ_B around 1.5 and 0.5, respectively, to

give acceptable classification performance.

3 Material

We use human nucleolus-associated proteins from public inventories as collected by Hinsby *et al.* (2006). After minor updates (including the proteins discovered as a result of their study) this set contains 879 proteins. We re-use the Hinsby *et al.* (2006) protein-protein interaction network which was constructed from BIND (Alfarano *et al.*, 2005) involving known human nucleolus proteins and their interaction partners (including matches of close orthologs). The network has in total 2607 vertices/proteins and 26401 first-order interactions. To test the proposed methods, the network is cleaned so that only vertices representing known nucleolar proteins with corresponding edges are retained. All interactions encoded by the diffusion kernel for presentation to KCCA thus only involve nucleolar proteins. The resulting 474 proteins (with 3974 interactions) constitute our final positive set. The set was not further reduced since it derives from a single genome.¹

A negative set was composed from non-nucleolar proteins in the Nuclear Protein Database (Dellaire *et al.*, 2003) and from a heavily filtered search of mammalian proteins in UniProt R51. UniProt proteins were required to have a non-ambiguous nuclear subcellular localisation with further intra-nuclear association explicitly stated, not including the nucleolus. We further cleaned the negative set by removing all proteins that were in the 879-nucleolus-associated sequence set (or homologs thereof). Finally, to prevent over-estimation of test accuracy, the negative set was reduced so that the remaining sequences had less than 30% similarity (removing any orthologs). The final negative 359 sequence set thus represents nuclear proteins with no experimentally confirmed association with the nucleolus. However, due to the

¹Only minimal redundancy was found. Our final positive set contains seven proteins above the default BlastClust similarity threshold.

inherent fluidity of nuclear proteins, the negative set may still contain proteins that are transiting through the nucleolus. The negative proteins were added to the protein interaction network as unconnected vertices.

4 Results

We use 5-fold cross-validation to provide estimates of expected accuracies for models when tested on truly novel proteins. Thus, in each “fold” 4/5 of data are used for training and 1/5 is set aside for testing. Within a cross-validation simulation, the procedure is repeated five times, so that each subset is a test set once. For a cross-validation simulation, we report the test accuracy for all proteins, but only when they are included as part of the test set. Importantly, the interactions involving any test protein are absent from the network used for training (this reduced the number of edges to about 2500 in each fold). In all cases, we quantify test accuracies when a cross-validation simulation is repeated ten times (with shuffled data).

4.1 Inferring protein-protein interaction

Each test protein was analysed to determine how well its KCCA induced representation correlated with known network components. We specifically looked at the ten protein “proxies” with the highest correlation and their interaction partners. As described previously, predicted interaction partners are weighted using the correlation with which they associate to the query protein. Each test protein is thus predicted to interact with a number of training proteins (the predicted edge is never in the training data). All test proteins were collectively used to predict a complete network. It should be noted that this is quite challenging since not only is protein-protein interaction data notoriously noisy but in several cases not all correct interactions are in the training data. To understand the ability of KCCA to identify correct inter-

actions, we introduce a correlation threshold, θ , which needs to be exceeded for an interaction to be included in the predicted network.

By initially including all associated interactions (bounded in numbers by those interactions present in the ten most highly correlated proxy proteins) and then gradually increasing θ , the recall and precision of the model can be studied. We intentionally avoid including a negative prediction score in our analysis due to the large number of true negatives in network inference.² Instead we focus on recall and precision, defined as $\frac{tp}{tp+fn}$ and $\frac{tp}{tp+fp}$, respectively. tp is the number of true positives, fp the number of false positives, and fn the number of false negatives. We choose six levels at which the recall and precision of each model configuration are monitored: the network (with all proteins) at circa 8000, 4000, 1000, 500, 100 and 50 predicted interactions. To summarise the models' ability of inference, we show an average recall at levels 8000, 4000 and 1000, and an average precision at levels 500, 100 and 50 interactions.

Overall, we found that network inference worked best when KCCA involved the Local Alignment kernel. However, attesting to the challenging test scenario, average recall was at best exceeding 0.10. Average precision varied greatly, but there were configurations when it approached 0.30. When a high θ was imposed, specific interactions were inferred with confidence.

An overview of network inference performance is provided in Table 1. The Local Alignment kernel performed best leading to inference with an average recall of 0.12 and an average precision of 0.27. By increasing the order of the Spectrum and Wildcard kernel from 3 to 4, the average precision increased from 0.10 to 0.19 and 0.16 (respectively), while average recall dropped slightly. However, with higher order we noted worse overall classification as outlined in the next section.

²With $474 + 359 = 833$ proteins, there are $(833^2 - 833)/2 = 346528$ possible binary (undirected) interactions of which only 3974 are positives. The lack of a meaningful true negative rate is problematic for standard ROC analysis.

---Table 1 about here---

An example (average-scoring) network predicted by the Local Alignment kernel and filtered to include about 50 of its strongest predictions is shown in Figure 3. A similar example is shown in Figure 4. In the networks we show putative functions as made available through the Nucleolar Proteome Database (Leung *et al.*, 2006). Gene Ontology terms for all UniProt proteins were checked for overrepresentation using BiNGO (Maere *et al.*, 2005).

When assessing accuracy of the predicted networks, it should be noted that the “gold standard” interaction data is sparse and noisy (von Mering *et al.*, 2002). Thus, the number of correct predictions is likely an underestimate.

---Figure 3 about here---

---Figure 4 about here---

4.2 Classifying nucleolar association

Turning to the first method for classifying nucleolar association (identifying the $n = 10$ most correlated proteins before weighting their respective class-memberships), we note that the regularisation settings are different from those that worked for network inference. $\lambda_A = 0.7$ and $\lambda_B = 0.1$ were seen in general to give best results. Notably, this particular setting rendered poor network inference performance. The classification results for the first method are summarised in Table 2. We use the correlation coefficient (CC) between experimentally confirmed association with the nucleolus and the prediction to illustrate the accuracy.

$$CC = \frac{tp \cdot tn - fp \cdot fn}{\sqrt{(tp + fn)(tp + fp)(tn + fp)(tn + fn)}} \quad (3)$$

where tn is the number of true negatives (tp , fp , and fn were introduced in Section 4.1).

---Table 2 about here---

As described previously, the second classification method relies solely on a sequence kernel and a support vector machine. The results for method two, which generally exceed those of method one, are summarised in Table 3. The method is similar to that of Lei and Dai (2005). They report a lower accuracy for nucleolar proteins (CC=0.37) but attempt the more challenging problem of distinguishing between six intra-nuclear compartments. Hinsby *et al.* (2006) report a correlation coefficient of 0.48 for a neural network that, similar to our second method, simply discriminates between nucleolar and non-nucleolar proteins. Their neural network takes as input features like protein disorder, iso-electric point and the presence of nuclear localisation signals. It should be noted that their negative set is not specified, preventing a more detailed comparison.

---Table 3 about here---

The third method uses a support vector machine that combines the KCCA produced vector and a sequence kernel. The results are summarised in Table 4. It should be noted that the method is only based on the sequence data but seen through the eyes of KCCA as well as a separate sequence kernel. The results firmly indicate that the combination of the support-vector machine and KCCA (tuned from network data) boosts classification performance beyond that of a method two.

---Table 4 about here---

To qualify the contribution of the protein-protein interaction data, we looked at all mis-classified nucleolar proteins as identified through ten repeats of a cross-validation simulation both of method two and of method three. The function, as annotated by the Nucleolar Proteome Database (Leung *et al.*, 2006), was used to group the proteins mistaken as non-nucleolar. The average number of proteins for each function is shown

in Table 5. A striking improvement of method 3 over the sequence-only method two is seen for ribosomal and RNA-related proteins, key functions of the nucleolus. As detailed in Table 5, if the “unknowns” are excepted, the total number of proteins in each group correlates well with the absolute improvement of method 3 (the Pearson correlation is 0.79) and with the total number of interactions that each protein is involved in (correlation is 0.66). Hence, core proteins of the nucleolar molecular aggregate benefit substantially from the integrative approach of method 3.

---Table 5 about here---

5 Conclusion

The kernel-based framework presented herein is able to tap the most recent genomics data resources to enable a true integrative, multi-level approach to systems biology. In particular, KCCA enables the construction of predictors which are based on ubiquitous genomic sequence data but leverages other knowledge sources during training.

With diffusion-based transport and molecular interaction as the primary means of compartmental association, intra-nuclear trafficking is modelled effectively using the kernel-based framework. We present a first attempt to both untangle nucleolar protein interactions but primarily to classify nucleolar association with accuracies that surpass baseline approaches.

The full known nucleolar protein-protein interaction network is inferred with an average recall of about 0.12 and an average precision of about 0.27. Higher precision is achieved when predictions are filtered according to KCCA correlation. In configurations of high precision we are able to identify small potential protein complexes but predictions are not without spurious cases. Cell-cycle/co-expression and other types of data may further assist in removing false positives introduced by interaction predictions (de Lichtenberg *et al.*, 2005).

The simpler problem of classifying novel proteins as associating with the nucleolus renders a correlation coefficient of about 0.54 (which corresponds to about 78% accuracy on our data set). Different optimal settings of KCCA regularisation parameters indicate a tension between accurate nucleolar association and protein-protein interaction. Nucleolar components are fluid and interact with several members that are non-nucleolar. Hence, accurate prediction of protein-protein interaction may thus introduce influences impeding the classification problem. For both network inference and classification we found the Local Alignment kernel to provide superior guidance for KCCA and the support vector machine, respectively.

The contribution of the interaction network for classification accuracy is seen in particular for core, ribosomal and RNA-related nucleolar proteins. An intriguing extension to the current study thus involves the inclusion of protein-RNA and protein-DNA interactions into the network encoded by the diffusion kernel, completing the repertoire of molecular binding partners of the nucleolus. Since the nucleolus disintegrates in the absence of transcriptional activity (at its site close to ribosomal genes), compartment association is likely dependent on ribosomal RNA, motivating their inclusion.

Acknowledgements

The nucleolar protein set and the protein-protein interaction set were collected and kindly made available by Anders Hinsby, Olof Karlberg and colleagues at the Technical University of Denmark. Lynne Davis helped implementing some of the algorithms. James Watson ran extensive simulations to determine suitable regularisation parameter values. The research was supported by the ARC Centre of Complex Systems. R.D.T. is supported by an NHMRC R. Douglas Wright Career Development Award.

References

- Akaho, S. (2001). A kernel method for canonical correlation analysis. In *Proceedings of the International Meeting of Psychometric Society – IMPS2001*.
- Alfarano, C., Andrade, C. E., Anthony, K., Bahroos, N., Bajec, M., Bantoft, K., Betel, D., Bobechko, B., Boutilier, K., Burgess, E., *et al.* (2005). The biomolecular interaction network database and related tools 2005 update. *Nucleic Acids Research*, **33**, D418–D424.
- Andersen, J. S., Lam, Y. W., Leung, A. K., Ong, S. E., Lyon, S. E., Lamond, A. I., and Mann, M. (2005). Nucleolar proteome dynamics. *Nature*, **433**, 77–83.
- Bach, F. R. and Jordan, M. I. (2002). Kernel independent component analysis. *Journal of Machine Learning Research*, **3**, 1–48.
- Bickmore, W. and Sutherland, H. (2002). Addressing protein localization within the nucleus. *The EMBO Journal*, **21**, 1248–1254.
- Carmo-Fonseca, M. (2002). The contribution of nuclear compartmentalization to gene regulation. *Cell*, **108**(4), 513–521.
- de Lichtenberg, U., Jensen, L. J., Brunak, S., and Bork, P. (2005). Dynamic complex formation during the yeast cell cycle. *Science*, **307**, 724–727.
- 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.
- Dundr, M. and Misteli, T. (2001). Functional architecture in the cell nucleus. *Biochem. J.*, **356**(2), 297–310.
- Geurts, P., Touleimat, N., Dutreix, M., and d’Alché Buc, F. (2007). Inferring biological networks with output kernel trees. *BMC Bioinformatics*, **8**(S4).
- Hinsby, A. M., Kiemer, L., Karlberg, E. O., Lage, K., Fausboll, A., Juncker, A. S., Andersen, J. S., Mann, M., and Brunak, S. (2006). A wiring of the human nucleolus. *Molecular Cell*, **22**, 285–295.
- Kato, T., Tsuda, K., and Asai, K. (2005). Selective integration of multiple biological data for supervised network inference. *Bioinformatics*, **21**(10), 2488–2495.
- Kondor, R. I. and Lafferty, J. (2002). Diffusion kernels on graphs and other discrete input. In *Proceedings of the International Conference on Machine Learning (ICML 2002)*, pages 315–322. Morgan Kaufmann Press.
- Lei, Z. and Dai, Y. (2005). An SVM-based system for predicting protein subnuclear localizations. *BMC Bioinformatics*, **6**, 291.
- Lei, Z. and Dai, Y. (2006). Assessing protein similarity with Gene Ontology and its use in subnuclear localization prediction. *BMC Bioinformatics*, **7**, 491.

- Leslie, C. and Kuang, R. (2004). Fast string kernels using inexact matching for protein sequences. *Journal of Machine Learning Research*, **5**, 1435–1455.
- Leung, A. K. L., Andersen, J. S., Mann, M., and Lamond, A. I. (2003). Bioinformatic analysis of the nucleolus. *Biochemical Journal*, **376**, 553–569.
- Leung, A. K. L., Trinkle-Mulcahy, L., Lam, Y. W., Andersen, J. S., Mann, M., and Lamond, A. I. (2006). NOPdb: Nucleolar Proteome Database. *Nucleic Acids Research*, **34**(S1), D218–220.
- Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in biological networks. *Bioinformatics*, **21**, 3448–3449.
- Misteli, T. (2001). Protein dynamics: Implications for nuclear architecture and gene expression. *Science*, **291**(5505), 843–847.
- 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.
- Schölkopf, B., Simard, P., Smola, A., and Vapnik, V. (1998). Prior knowledge in support vector kernels. In M. Jordan, M. Kearns, and S. Solla, editors, *Advances in Neural Information Processing Systems 10*, pages 640–646, Cambridge, MA. MIT Press.
- Stein, G. S., van Wijnen, A. J., Stein, J. L., Lian, J. B., Montecino, M., Choi, J., Zaidi, K., and Javed, A. (2000). Intranuclear trafficking of transcription factors: implications for biological control. *J Cell Sci*, **113**(14), 2527–2533.
- Vapnik, V. (1998). *Statistical Learning Theory*. Wiley, New York.
- Vert, J.-P. and Kanehisa, M. (2003). Graph-driven features extraction from microarray data using diffusion kernels and kernel cca. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, pages 1425–1432. MIT Press.
- von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**, 399–403.
- Yamanishi, Y., Vert, J.-P., and Kanehisa, M. (2004). Protein network inference from multiple genomic data: a supervised approach. *Bioinformatics*, **20**, i363–i370.

Recall	LA(0.1)	WC(3,1)	MM(3,1)	SP(3)
8000	0.18 ± 0.02	0.18 ± 0.01	0.18 ± 0.01	0.16 ± 0.01
4000	0.13 ± 0.01	0.12 ± 0.01	0.10 ± 0.01	0.10 ± 0.01
1000	0.04 ± 0.01	0.04 ± 0.00	0.03 ± 0.01	0.03 ± 0.00
500	0.03 ± 0.01	0.02 ± 0.00	0.02 ± 0.01	0.02 ± 0.00
Precision	LA(0.1)	WC(3,1)	MM(3,1)	SP(3)
1000	0.18 ± 0.03	0.09 ± 0.01	0.13 ± 0.03	0.11 ± 0.01
500	0.22 ± 0.05	0.12 ± 0.02	0.16 ± 0.06	0.12 ± 0.03
100	0.33 ± 0.06	0.14 ± 0.03	0.12 ± 0.13	0.12 ± 0.06
50	0.26 ± 0.07	0.16 ± 0.02	0.04 ± 0.06	0.09 ± 0.07

Table 1: Protein-protein interaction network inference performance for different sequence kernels (LA is the Local Alignment, WC is the Wildcard, MM is the Mismatch and SP is the Spectrum kernel). All results are based on five-fold cross-validation, repeated ten times with different data set divisions (standard deviations are reported after ‘ \pm ’). $\lambda_A = 0.9$ and $\lambda_B = 1.3$ as determined from parameter sweeps. Recall indicates the ability to recover the target network. Precision indicates the correctness of predicted interactions.

Kernel/config	Correlation coefficient
LA(0.1)	0.42 ± 0.02
WC(3,1)	0.36 ± 0.03
MM(3,1)	0.32 ± 0.02
SP(3)	0.36 ± 0.02
WC(3,1)*	0.37 ± 0.02
MM(3,1)*	0.36 ± 0.02
WC(3,1)* $n = 20$	0.40 ± 0.01
MM(3,1)* $n = 20$	0.38 ± 0.02

Table 2: Classification performance measured as the correlation between experimentally confirmed association with the nucleolus and method one’s prediction. Results are reported for different sequence kernels. All results are based on five-fold cross-validation, repeated ten times with different data set divisions (standard deviations are reported after ‘ \pm ’). Default $n = 10$, $\lambda_A = 0.7$ and $\lambda_B = 0.1$ as determined from parameter grid search. Alternate configurations marked with asterisk were found by scanning all $\lambda_A = \lambda_B$ up to 10.0, showing $\lambda_A = \lambda_B = 3.0$ as best.

Kernel/config	Correlation coefficient
LA(0.1)	0.48 ± 0.01
WC(3,1)	0.46 ± 0.02
MM(3,1)	0.47 ± 0.01
SP(4)	0.41 ± 0.02
SP(3)	0.47 ± 0.01
SP(2)	0.47 ± 0.02
SP(1)	0.45 ± 0.01

Table 3: Classification performance measured as the correlation between experimentally confirmed association with the nucleolus and method two’s prediction. Results are reported for different sequence kernels. All results are based on five-fold cross-validation, repeated ten times with different data set divisions (standard deviations are reported after ‘ \pm ’). Default SVM regularisation parameter $C = 1.0$ as determined from preliminary trials.

KCCA A -kernel	Direct kernel	λ_1, λ_2	Correlation coefficient
LA(0.1)	LA(0.1)	1.3, 0.7	0.54 ± 0.01
LA(0.1)	WC(3,1)	1.3, 0.7	0.54 ± 0.01
WC(3,1)	LA(0.1)	1.8, 0.2	0.53 ± 0.01
WC(0.1)	WC(3,1)	1.8, 0.2	0.51 ± 0.01

Table 4: Classification performance measured as the correlation between experimentally confirmed association with the nucleolus and method one’s prediction. Results are reported for different sequence kernels. All results are based on five-fold cross-validation, repeated ten times with different data set divisions (standard deviations are reported after ‘ \pm ’). Default $n = 10$, $C = 5$, λ_A and λ_B as determined from parameter grid search.

Function	Method 2	Method 3	Vertices	Edges
Apoptosis	0.0	0.0	1	2
Cell cycle related factor	4.7	4.1	14	62
Chaperone	0.1	0.1	10	107
Chromatin related factor	1.6	1.2	8	39
Contaminant	0.7	0.1	11	68
Cytoskeleton	0.0	0.6	8	82
DNA binding protein	1.6	1.5	9	33
DNA helicase	0.1	0.0	4	71
DNA methyltransferase	1.0	0.7	1	3
DNA repair	2.2	1.4	6	23
DNA replication	1.4	1.7	15	112
Exonuclease mRNA	0.6	0.1	1	1
G-protein	0.6	0.0	1	53
GTP binding protein	0.0	0.0	4	139
Histone modifying factor	0.5	0.3	4	18
Importin/exportin	1.0	0.9	3	13
Intermediate filaments	1.0	1.0	2	24
Kinase/phosphatase	1.7	2.2	9	194
Lamin	1.0	1.0	1	2
Other translation factors	3.1	2.3	15	190
Proliferation	1.0	1.0	1	76
RNA binding protein	3.6	0.9	13	213
RNA helicase	2.0	0.9	20	548
RNA modifying enzymes	5.5	1.9	34	1031
RNA polymerase	2.5	1.4	9	123
Ribosomal proteins	4.1	0.5	34	199
Splicing related factor	6.5	4.5	23	340
Transcription factor	6.8	5.2	19	268
Ubiquitin related protein	2.1	1.0	6	125
WD-repeat protein	1.7	1.6	5	173
hnRNP	1.6	2.1	7	56
p53 activating	0.0	0.0	1	1
Unknown	34.6	32.7	175	3559

Table 5: The average number of mis-classified nucleolar proteins (of 474 proteins) grouped according to their function. The first column represents errors made by method two which uses only sequence data. The second column represents errors made by method three which is trained on both sequence and protein-protein interaction data. The third column (“Vertices”) indicates the total number of proteins in the functional group and the fourth column (“Edges”) shows the total number of interactions that emanate from each protein in the group.

Figure 1: Above: A protein-protein interaction network with three protein complexes. The protein accessions are NP_005024 (1), Q01780 (2), Q13868 (3), Q7Z481 (4), Q9NPD3 (5), Q9NQT5 (6), Q9Y3B2 (7), NP_478126 (8), Q15024 (9), Q96B26 (10), Q9NQT4 (11), P18887 (12), Q9P1V2 (13), P06746 (14), P49916 (15), P33993 (16) and P49736 (17). The indices apply to the kernel matrices. The network shown is a subset of the full nucleolar protein-protein interaction network. Below: The (normalised) diffusion kernel matrix for the network (cells with large values are shown as white, small values as black). The diffusion bandwidth is 0.5.

Figure 2: Above: The normalised wildcard kernel matrix. Below: The normalised local alignment kernel matrix. Both matrices are based on the 17 proteins in three nucleolar complexes (cells with large values are shown as white, small values as black). Sample indices are the same as in Figure 1. The wildcard kernel uses a pattern length of 3 with a single “wildcard match” accepted. The local alignment kernel uses $\beta = 0.1$ (balancing impact of non-optimal alignments).

Figure 3: An example predicted network with 48 interactions filtered according to strength of correlation (of which 17 are correct according to the target protein-protein interaction network, marked in bold). Each vertex is labelled with the protein's accession id and putative function. The network was inferred using KCCA and the Local Alignment/Diffusion kernel with λ -values 0.9 and 1.3, respectively. Specific notes: P02570 is an Actin-protein, O15142, O15144, O15145 and O15143 are actin-like proteins acting in a known complex. P12004 is a proliferating cell nuclear antigen (PCNA) member. PCNA helps hold DNA polymerase delta to DNA and has roles important for DNA repair and synthesis. DNA replication is highly over-represented ($p < 10^{-9}$) according to gene GO-terms/BiNGO.

Figure 4: An example predicted network with 47 interactions filtered according to strength of correlation (of which 28 are correct according to the target protein-protein interaction network, marked in bold). Each vertex is labelled with the protein's accession id and putative function. The network was inferred using KCCA and the Local Alignment/Diffusion kernel with λ -values 0.9 and 0.1, respectively. Specific notes: The predicted aggregate has associations with the 60S Ribosome subunit. 35S primary transcript processing, ribosome biogenesis, and rRNA processing all come up as overrepresented GO-terms in BiNGO ($p < 10^{-16}$).