PREDICTING NUCLEOLAR PROTEINS USING SUPPORT-VECTOR MACHINES

MIKAEL BODÉN

*ARC Centre of Excellence in Bioinformatics, Institute for Molecular Bioscience, and
School of Information Technology and Electrical Engineering
The University of Queensland, QLD 4072, Australia
E-mail: m.boden@uq.edu.au*

The intra-nuclear organisation of proteins is based on possibly transient interactions with morphologically defined compartments like the nucleolus. The fluidity of trafficking challenges the development of models that accurately identify compartment membership for novel proteins. A growing inventory of nucleolar proteins is here used to train a support-vector machine to recognise sequence features that allow the automatic assignment of compartment membership. We explore a range of sequence-kernels and find that while some success is achieved with a profile-based local alignment kernel, the problem is ill-suited to a standard compartment-classification approach.

Keywords: Nucleolus, support-vector machine, intra-nuclear protein localisation, kernel

1. Introduction

By virtue of its architecture, the cell nucleus not only encloses the genetic material but also controls its expression. Recent discoveries have exposed morphologically well-defined compartments with which proteins and RNA associate.[1,2] This paper uses emerging experimental data to develop a basic predictive model of intra-nuclear protein association.

Similar to cytoplasmic organelles, intra-nuclear compartments seem to specialize in particular functions (like ribosomal RNA synthesis, spliceosome recycling and chromatin remodeling). However, intra-nuclear compartments are not membrane-bound and thus employ different principles to sustain their functional integrity. Indeed, compartments are in perpetual flux, with some proteins and RNA stably associated and others just transiently binding before they move on to another compartment. Proteins and RNA are trafficked by non-directed, passive diffusion and association with a compartment is based on molecular interactions with its residents.[1,2] The largest compartment inside the nucleus is the nucleolus. With functions primarily related to ribosomal biogenesis, the nucleolus is conveniently located at sites of ribosomal genes. Apart from being involved in producing ribosomes, examples of nucleolar functions include the maturation of tRNA and snRNA of the spliceosome, pre-assembly of the signal recognition particle and the sequestration of several important regulatory proteins.[3]

Recent efforts using mass spectrometry have resulted in the identification of a substantial number of nucleolar proteins in human cells.[4] With the view that proteins are only transiently associated with one or more compartments, we ask if we can build a classifier that is able to distinguish between proteins with nucleolar association from those without. Specifically, a growing protein inventory is leveraged using state-of-the-art machine learning algorithms–support-vector machines equipped with sequence kernels. This paper develops an appropriate data set, and a sequence data-driven model. The model is evaluated on its ability to capture in terms of sequence features the possibly loose association of proteins with the nucleolus.

2. Background

Analysis has shown that there seems to be no single feature that allows the automatic sorting of proteins into nuclear compartments.[5] Several characteristics, like iso-electric point, molecular weight, and amino acid and domain composition may need to be used in conjunction to accurately assign their compartmental association.[5] The nucleolus has the largest number of known proteins, but there appears to be few generic motifs shared by its residents, the so-called DEAD-box helicase and the WD40 repeat being two notable exceptions each occurring in about 6% of known members.[5]

Using the Nuclear Protein Database,[6] Lei and Dai[7,8] 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 to a classifier. 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 (via a BLAST search). Additionally, a separate support-vector machine is trained to map the sequence to one of the six classes. Notably, inclusion of the GO term module elevates overall performance considerably (the correlation coefficient for nucleolus improves from 0.37 to 0.66). However, the GO terms (a) include specific annotations of localisation and (b) need to be known in advance.

Hinsby et al.[3] 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 identified prioritised nucleolar protein complexes and subsequently eleven novel nucleolar proteins (by targeted search for 55 candidates in the raw mass spectrometry data). The approach indicates the potential of assigning intra-nuclear compartment membership in terms of interactions with residents rather than possibly elusive compartment-unifying features.

3. Methods

3.1. *Data set*

We re-use the data set of Hinsby et al.,[3] sourced primarily from the Nucleolar Proteome Database (NOPdb[9]), then adding the eleven novel proteins from Hinsby et al.'s study, resulting in 879 human nucleolus-localised proteins. We further performed redundancy reduction using BlastClust ensuring that only 30% sequence similarity was present in the remaining set of 767 positives. This set consists of proteins which are either stable or transient residents of the nucleolus. Importantly, they could also be present in other locations to varying degrees.

Preliminary investigations which did not employ a negative training set were unsuccessful. More specifically, we used one-class support-vector machines to generate a decision function that included only all positives. Test performance on known negatives clearly indicated the need for pursuing a full discriminative approach. Thus, a negative, non-nucleolar protein set was devised from two sources: the Nuclear Protein Databank[6] and UniProt R51–restricted to mammalian proteins. NPD-extracted proteins had one or more intra-nuclear compartments assigned, not including the nucleolus. UniProt proteins were similarly required to have a non-ambiguous nuclear subcellular localisation with further intra-nuclear association explictly stated, not including the nucleolus. We further cleaned the negative set by removing all proteins that were in the original positive 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.

The final negative 359-sequence set thus represents nuclear proteins with no experimentally confirmed association with the nucleolus. However, due to the inherent fluidity of nuclear proteins, the negative set may still contain proteins that are transiting through the nucleolus. It should be noted that the final data sets differ from the sets used by Lei and Dai who removed any protein not exclusively associated with one of the six compartments. Additionally, 35 nucleolar proteins were found in the original 879-set that were incorrectly assigned exclusively to a non-nucleolar compartment in their study.

3.2. *Model*

Support-vector machines (SVMs[10]) are trained to discriminate between positive and negative samples, i.e. to generate a decision function

$$f(\mathbf{x}) = \sum_{i=1}^{n} y_i \alpha_i K(\mathbf{x_i}, \mathbf{x}) + b \qquad (1)$$

where $y_i \in \{-1, +1\}$ is the target class for sample $i \in \{1, ..., n\}$, $\mathbf{x_i}$ is the ith sample, α_i is the ith Lagrange multiplier and b is a threshold. All multipliers and the threshold are tuned by training the SVM.

To determine the Lagrange multipliers, Platt's sequential minimal optimization[11] with convergence improvements[12] is used. Note that only multipliers directly associated with samples on the margin separating positives from negatives are non-zero (these samples are known as the support-vectors). Models based on support-vector machines have previously garnered success for classifying cytoplasmic protein compartmentalisation.[13–16]

Due to the graded membership of intra-nuclear compartments, the SVM output is converted to a probabilistic output, using a sigmoid function

$$p(\mathbf{x}) = \frac{1}{1 + e^{A \cdot f(\mathbf{x}) + B}} \tag{2}$$

where A and B are estimated by minimizing the negative log-likelihood from training samples.[17] The training data assigned to the model is divided internally so that approximately 4/5 is used for tuning the support-vector machine, and 1/5 for tuning the sigmoid function.

A number of sequence-based kernels have been developed recently, primarily targeted to protein classification problems. We evaluate the performance of the Spectrum kernel,[18] the Mismatch kernel,[19] Wildcard kernel,[19] the Local Alignment (LA) kernel[20] and a profile-based Local Alignment kernel, each replacing $K(\cdot,\cdot)$ in Equation 1.

We refer the reader to the literature for detailed information regarding the kernels. Essentially, spectrum-based kernels (including the Mismatch and Wildcard kernels) are based on the sharing of short sequence seqments (of length k, with provision of minor differences, m is the allowed number of "mismatches" in the Mismatch kernel, x is the number of "wildcard" symbols in the Wildcard kernel).[19]

The Local Alignment kernel compares two sequences by exploring their alignments.[20] We explore some details of the Local Alignment kernel to describe the only novel kernel in this paper–the Profile Local Alignment kernel.

An alignment between two sequences is quantified using an amino acid substitution matrix, S, and a gap penalty setting, g. A further parameter, β, controls the contribution of non-optimal alignments to the final score. Let $\Pi(\mathbf{x_1}, \mathbf{x_2})$ be the set of all possible alignments between sequences $\mathbf{x_1}$ and $\mathbf{x_2}$. The kernel can be expressed in terms of alignment-specific scores, $\varsigma_{S,g}$ (for details of this function see[20]).

$$K_\beta^{LA}(\mathbf{x_1}, \mathbf{x_2}) = \sum_{\pi \in \Pi(\mathbf{x_1}, \mathbf{x_2})} exp(\beta \varsigma_{S,g}(\mathbf{x_1}, \mathbf{x_2}, \pi)) \tag{3}$$

When the Local Alignment kernel is used herein, S is the BLOSUM62 matrix.

Evidence is mounting that so-called position-specific substitution matrices (PSSMs; a.k.a. "profiles") disclose important evolutionary information tied to each residue.[21,22] We adapt the alignment-specific function, ς, in the Local Alignment kernel to use such substitution scores generated by PSI-Blast (max three iterations, E-value threshold is 0.001, using Genbank's non-redundant protein set) in place of the generic substitution matrix, S. Specifically, we define the substitution score as the average of the PSSM-entries for the two sequences (where the entry coordinates are determined from the sequence position of one sequence and the symbol of the other).

Table 1. Accuracy of classification for different kernel settings when the output cut-off is set to 0.5. Mean correlation coefficient on test data in 10-fold crossvalidation, repeated 10 times, is shown (1.0 indicates ideal agreement, 0.0 indicates chance agreement with target data). The standard deviation is provided for each configuration after ±.

Kernel	**Parameters**	**Correlation coefficient**
Spectrum	$k = 3$	0.340 ± 0.016
Wildcard	$k = 3,\ x = 1$	0.391 ± 0.012
Wildcard	$k = 4,\ x = 1$	0.388 ± 0.013
Mismatch	$k = 3,\ m = 1$	0.382 ± 0.015
Mismatch	$k = 4,\ m = 1$	0.420 ± 0.017
Local Alignment	$\beta = 0.1$	0.399 ± 0.012
Profile Local Alignment	$\beta = 0.1$	0.447 ± 0.017

4. Results

Models are trained and tested using 10-fold crossvalidation. Essentially, the available data is first partitioned into ten evenly sized sub-sets. Second, ten models are trained on 9 of the ten sub-sets, each sub-set combination chosen so that it is unique. Third, each of the ten models is tested only on their respective remaining sub-set. Note that no model is trained on any of their test samples, and each of the original samples is used as a test sample by exactly one model. Finally, the test results are collated and the whole crossvalidation procedure is repeated ten times to establish variance in prediction accuracy.

All kernels are normalised, i.e. kernel values are adjusted such that the diagonal of the kernel matrix is 1.0. Due to substantive computation requirements, only a few kernel parameters were trialled but care was exercised to explore the configurations most successful in the literature.

Support-vector machines require the manual setting of regularisation parameters (C-values). Preliminary parameter-sweeps with two C-values (one for the positive and one for the negative set) identified that when they exceed 1.0 the support-vector machine generalised stably for all kernels. C-values were thus fixed at 1.0 throughout.

We use the correlation coefficient (CC) between experimentally confirmed association with the nucleolus and the prediction to illustrate the accuracy.

$$CC = \frac{tp \cdot tn - fp \cdot fn}{\sqrt{(tp + fn)(tp + fp)(tn + fp)(tn + fn)}} \tag{4}$$

where tp, tn, fp and fn is the number of true positives, true negatives, false positives and false negatives, respectively.

The classification of proteins as nucleolar-associated (or not) reached 77% accuracy on our data set with a SVM equipped by the Profile Local Alignment kernel. This corresponds to a correlation of $CC = 0.447$ (± 0.017) between observed and predicted nucleolar association. All classification results when using the default output cut-off at 0.5 are presented in Table 1.

To further illustrate the accuracy we generated ROC curves for the SVMs with the Profile Local Alignment kernel and the Mismatch kernel (see Figure 1). That is, by varying the threshold which needs to be exceeded by the probabilistic output, the sensitivity and specificity of the model is monitored.

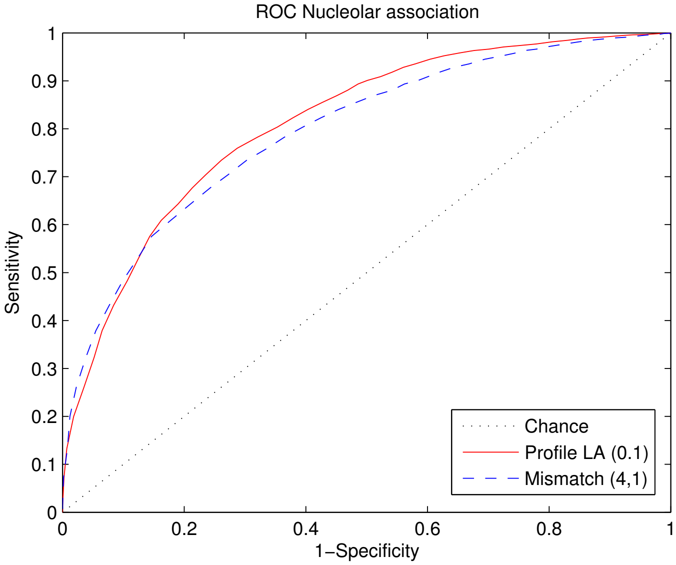

Fig. 1. ROC curves illustrating the change in sensitivity as a function of specificity. The area under the ROC is 0.811 for the Profile LA kernel ($\beta = 0.1$) and 0.794 for the Mismatch kernel ($k = 4$, $m = 1$). Maximum correlation coefficient 0.451 of the Profile LA SVM is seen at an output threshold of 0.66 (sensitivity=0.71, specificity=0.76). Sensitivity is defined as $tp/(tp + fn)$ and specificity as $tn/(tn + fp)$.

The probabilistic output has the potential of indicating the certainty of the prediction. We computed the mean output for the four classification outcomes using a 0.5 cut-off (again over 10 runs using our best configuration, i.e. over $(767+359) \cdot 10$ test samples). (a) A true positive is 0.81 (±0.12), (b) a false positive is 0.71 (±0.12), (c) a true negative is 0.26 (±0.13) and (d) a false negative is 0.34 (±0.12). Hence, it is reasonable to regard a prediction closer to the cut-off as uncertain.

In the absence of known motifs clearly identifying nucleolar association, we attempted to characterise the basis of generalisation of the best predictive model by qualifying the mistakes made by it.

Over all ten runs, we collated all proteins mistakenly predicted to be nucleolar. These false positives were divided into their location as assigned by the Nuclear Protein Database[6] and as used as training data by Lei and Dai.[7] Available assignments are shown in Table 2. The reader is reminded that this data set has limited coverage, thus we present ratios based on available data. The mistakes are seemingly distributed evenly between alternative intra-nuclear locations. Noteworthy, we discovered one protein (O95347) that was consistently misclassified as nucleolar. O95347 is indeed nucleolar according to NPD but associated with Chromatin in UniProt.

Table 2. Number of proteins falsely classified as nucleolar and their location according to the Nuclear Protein Database as used by Lei and Dai. Average counts (of 359 possible) are shown over 10 repeats of 10-fold crossvalidation tests. The "absolute" percentage of a mistaken location refers to the location-count over the total number of false positives. The "relative" percentage refers to the location-count relative the number of proteins known in each location in Lei and Dai's data set (assuming the distribution of proteins is uniform).

Location	**Proteins (count)**	**% (absolute)**	**% (relative)**
Chromatin	26.4	15	21
Lamina	30.4	17	27
Nucleolus	1.0	1	0
Nucleoplasm	25.4	15	17
PML	14.8	8	19
Speckles	17.9	10	16
Unknown	58.6	34	

We similarly collated all proteins that were incorrectly predicted to not associate with the nucleolus. The false negatives were cross-checked by identifying their function according to the Nucleolar Proteome Database.[9] Hence, the tabulation seen in Table 3 illustrates functions commonly confused with alternative locations. Not surprising, beside the "unknowns", at the top of the list there are functions that relate to alternative compartments rather than being uniquely nucleolar, e.g speckles are associated with both splicing and transcription related factors[2] and the nuclear lamina consists mainly of filament proteins, lamins.

On average a model in one fold of a cross-validation run is trained on about 1000 samples. Of these, about 600 were usually selected to be support-vectors, ultimately defining the model's decision boundary. To further qualify the nature of subscribed generalisation, about 10% of all support-vectors of one model were analysed using a kernelised hierarchical cluster analysis (using normalised Profile Local Alignment kernel and average-linkage). The cluster dendrogram is shown in Figure 2. Each support-vector is labelled with its target label (Pos=Nucleolar or Neg=Other locations), function as determined from the Nucleolar Proteome Database or location as used by Lei and Dai. Proteins without functional annotation or location were excluded. Functional groups are visible (e.g. splicing/transcription, chromatin, lamina/cytoskeleton) further indicating that generalisation is based on protein function rather than intra-nuclear location.

5. Conclusion

We develop a model that is able to predict nucleolar association of proteins from their sequence. A support-vector machine fitted with a profile-based adaptation of

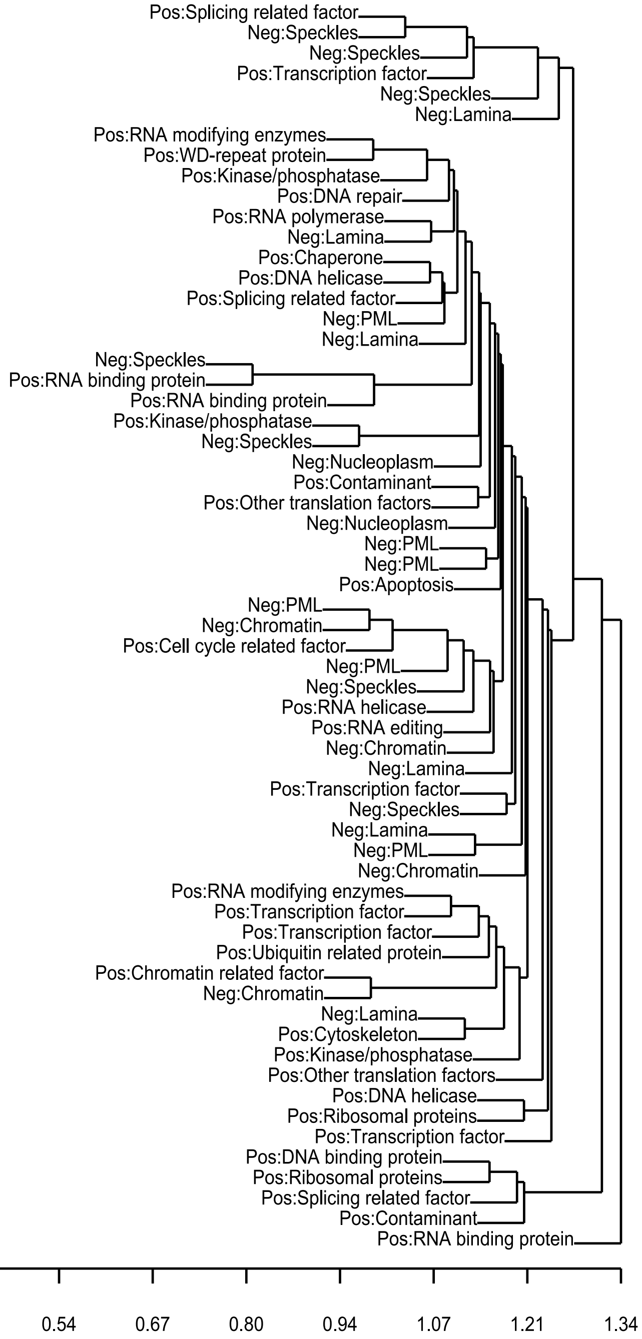

Fig. 2. A cluster dendrogram illustrating the organisation of the support-vectors implemented by one arbitrarily selected model. The distance in feature-space is indicated by the horisontal axis. Several clusters exemplify the mixing of location but coherence of function, e.g. on top nucleolar splicing and transcription factors are mixed with speckles-native proteins (the primary site for such functions).

Table 3. Number of proteins falsely predicted as non-nucleolar and their function according to the Nucleolar Proteome Database. Average counts (out of 767 positives) are shown over 10 repeats of 10-fold crossvalidation tests.

Function	**Proteins (count)**
Function unknown	49.9
Cell cycle related factor	4.7
Transcription factor	3.9
Splicing related factor	3.8
Ubiquitin related protein	2.2
DNA binding protein	1.8
Lamina	1.8
Kinase/phosphatase	1.7
WD-repeat protein	1.7
Contaminant	1.6
RNA binding protein	1.5
RNA modifying enzymes	1.4
p53 activating	1.3
DNA repair	1.0
Intermediate filaments	1.0
RNA polymerase	1.0
Chromatin related factor	0.5
Chaperone	0.4
Other translation factors	0.4
DNA methyltransferase	0.1
Exonuclease mRNA	0.1

the Local Alignment kernel and a probabilistic output achieves a correlation coefficient of about 0.45 (or 77% on our specific data set). It is difficult to directly compare this result with Lei and Dai's work since their ensemble predictor distinguishes between six classes as well as using differently scoped training and test data. Their SVM-only model has a lower correlation coefficient, but their GO term model (which requires the prior identification of such terms, some of which are explicitly concerned with location) exceeds the accuracy presented herein.

Compartmentalisation of proteins inside the nucleus is fluid and categorically discriminating between such compartments may thus be objectionable. To alleviate issues with multiple localisations, positive data used for model-tuning did not exclude proteins for which additional compartments were known. Moreover, the model presented here incorporates a probabilistic output which allows graded membership to be reflected.

Analysis shows that false positive predictions are drawn evenly from other intra-nuclear compartments. Conversely, nucleolar proteins not recognised as such are sometimes involved in functions also associated with alternative locations, suggesting that generalisation is based on functional features. Compartment-specific features are thus largely eluding an approach that has garnered success for cytoplasmic localisation, suggesting that to combat intra-nuclear trafficking we may need to reconsider model designs.

Acknowledgments

This work was supported by the ARC Centre of Complex Systems. Lynne Davis and Johnson Shih both contributed to this paper by implementing some of the kernels.

References

1. T. Misteli, *Science* **291**, 843 (2001).
2. K. E. Handwerger and J. G. Gall, *Trends in Cell Biology* **16**, 19 (2006).
3. A. M. Hinsby, L. Kiemer, E. O. Karlberg, K. Lage, A. Fausboll, A. S. Juncker, J. S. Andersen, M. Mann and S. Brunak, *Molecular Cell* **22**, 285 (2006).
4. J. S. Andersen, Y. W. Lam, A. K. Leung, S. E. Ong, S. E. Lyon, A. I. Lamond and M. Mann, *Nature* **433**, 77 (2005).
5. W. Bickmore and H. Sutherland, *The EMBO Journal* **21**, 1248 (2002).
6. G. Dellaire, R. Farrall and W. Bickmore, *Nucl. Acids Res.* **31**, 328 (2003).
7. Z. Lei and Y. Dai, *BMC Bioinformatics* **6**, p. 291 (2005).
8. Z. Lei and Y. Dai, *BMC Bioinformatics* **7**, p. 491 (2006).
9. A. K. L. Leung, L. Trinkle-Mulcahy, Y. W. Lam, J. S. Andersen, M. Mann and A. I. Lamond, *Nucleic Acids Research* **34**, D218 (2006).
10. V. Vapnik, *Statistical Learning Theory* (Wiley, New York, 1998).
11. J. Platt, Fast training of support vector machines using sequential minimal optimization, in *Advances in Kernel Methods–Suport Vector Learning*, eds. B. Schölkopf, C. J. C. Burgess and A. J. Smola (MIT Press, Cambridge, MA, 1999) pp. 185–208.
12. S. S. Keerthi, S. K. Shevade, C. Bhattacharyya and K. R. K. Murthy, *Neural Computation* **13**, 637 (2001).
13. V. Atalay and R. Cetin-Atalay, *Bioinformatics* **21**, 1429 (2005).
14. D. Sarda, G. Chua, K.-B. Li and A. Krishnan, *BMC Bioinformatics* **6**, p. 152 (2005).
15. A. Garg, M. Bhasin and G. P. S. Raghava, *J. Biol. Chem.* **280**, 14427 (2005).
16. A. Pierleoni, P. L. Martelli, P. Fariselli and R. Casadio, *Bioinformatics* **22**, e408 (2006).
17. J. Platt, Probabilistic outputs for support vector machines and comparison to regularized likelihood methods, in *Advances in Large Margin Classifiers*, eds. A. Smola, P. Bartlett, B. Schölkopf and D. Schuurmans (MIT Press, Cambridge, MA, 2000)
18. C. Leslie, E. Eskin and W. S. Grundy, The spectrum kernel: A string kernel for svm protein classification, in *Proceedings of the Pacific Symposium on Biocomputing*, eds. R. B. Altman, A. K. Dunker, L. Hunter, K. Lauerdale and T. E. Klein (World Scientific, 2002).
19. C. Leslie and R. Kuang, *Journal of Machine Learning Research* **5**, 1435 (2004).
20. H. Saigo, J.-P. Vert, N. Ueda and T. Akutsu, *Bioinformatics* **20**, 1682 (2004).
21. H. Rangwala and G. Karypis, *Bioinformatics* **21**, 4239 (2005).
22. R. Kuang, E. Ie, K. Wang, K. Wang, M. Siddiqi, Y. Freund and C. Leslie, *Journal of Bioinformatics and Computational Biology* **3**, 527 (2005).