

Chapter 4

Discriminating nucleolar proteins from nuclear proteins: is it possible?

4.1 Introduction

In Lokum (Chapter 3), I tried to predict the conventional eukaryotic protein localisation categories which usually fall into one of the general localisation groups of cell organelles, cell membrane or extracellular space. Here, I investigate the possibility of fine tuning some of these predictions by trying to predict sub-organelle categories. As an example, I consider nuclear proteins, and try to classify proteins in this category under two labels: nuclear and nucleolar.

Proteins destined to the nucleus have to pass through the nuclear pores (Figure 4.1¹). Nuclear pores could be imagined as holes piercing the impenetrable, hard nuclear envelope which, unlike the ER or plasma membrane, does not permit proteins to cross the membrane of the organelle directly regardless of whether

¹The image, originally designed by Mike Jones (<http://en.wikipedia.org/wiki/User:Adenosine>) has been reproduced here under the “Attribution-Share Alike 2.5 Generic” license of Creative Commons.

they contain membrane spanning regions. This makes the translocation of nuclear proteins different from secretory pathway proteins, including that they do not contain any cleavable targeting signals. Nuclear localisation signals (NLS), which mediate the import of proteins into the nucleus, could be anywhere on the sequence, unlike the C-terminal ER retention signal (see Section 3.3.1.2 and Figure 3.8), for instance. They comprise short sequences of basic amino acids like Arginine (R) and Lysine (K) (see Figures 3.4h-i and A.1), and form short binding sites for recognition by other molecules. In 1986, Goldfarb *et al.* showed that mutations in the NLSs can impair nuclear localisation, but also, non-nuclear proteins can be targeted into the nucleus if artificial NLSs were added to them.

Previously, other subnuclear localisation compartments have been proposed for where RNA splicing related proteins (“nuclear speckles”) (Li & Bingham, 1991) accumulate, and also for small nuclear ribonucleoprotein (snRNP) components (“foci”) (Chang & Lin, 2001), but the major and most studied subnuclear compartment is the nucleolus. There is experimental evidence suggesting a sequence-dependent targeting into the nucleolus by means of Nucleolar Localisation Signals (NOSs) (Dang & Lee, 1989) which are similar in composition to NLSs. Because nucleolar proteins have to first pass through the nuclear pores just like any other nuclear proteins, it is quite reasonable to expect them to have similar sort of signals that mediate their passages. Furthermore, having no membrane around the nucleoli may suggest that localisation in nucleoli could actually be achieved through mainly molecular binding. In fact, in an experimental study some nucleolar proteins in mouse have been reported to carry only an NLS but no identifiable NOS (Maeda *et al.*, 1992). Therefore in addition to the presence

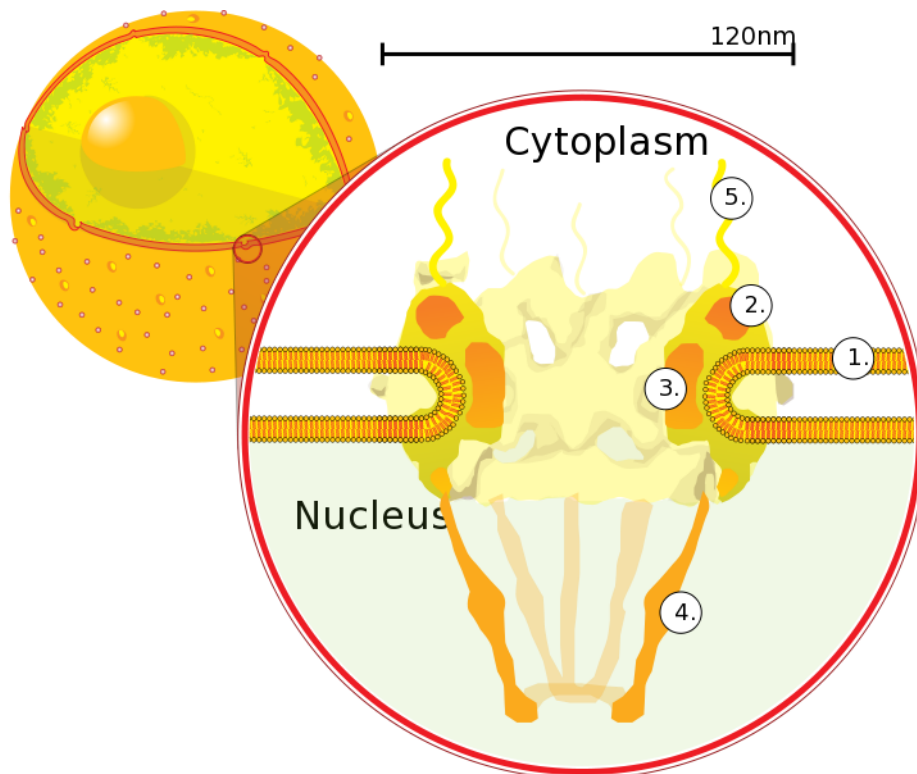


Figure 4.1: **Nuclear pore.** This schematic representation shows the nucleus, its nuclear envelope and a cross-section view of nuclear pores. Nuclear envelope is made of double membranes enclosing the genetic material in eukaryotic cells. Nuclear pores, crossing the nuclear envelope, allow water-soluble molecules to cross the nuclear envelope. Labels shown represent: 1 - Nuclear Envelope, 2 - Outer Ring, 3 - Spokes, 4 - Basket, and 5 - Filaments.

of NOSs, general protein properties such as amino acid composition could be important in nucleolar localisation.

The prediction of nuclear proteins is important because there are a lot of nuclear proteins in the cell, and difficult because the NLSs vary in sequence (Cokol *et al.*, 2000) and do not have specific positions. Prediction of nuclear proteins has probably begun with the multi-class localisation predictor PSORT (Nakai & Horton, 1999) which is based on many “if-then” type rules that comprise many biological features including discovered and known localisation signals (for a comparison of PSORT with Lokum see page 97). One of the more recent nuclear sequence prediction methods is PredictNLS (Cokol *et al.*, 2000). It predicts nuclear proteins by extrapolating from known NLSs which are listed in a specific database called NLSdb (Nair *et al.*, 2003). Initially, NLSdb had 114 experimentally determined NLSs that were obtained through an extensive literature search, but using ‘in silico mutagenesis’ this set was extended to 308 experimental and potential NLSs. PredictNLS is now part of a more general classifier, LOCTarget (Nair & Rost, 2004) that uses 4 specialised predictor programs: apart from NLSdb matches, it uses sequence homology (LOChom), SWISS-PROT keywords that are strongly correlated with localisation (LOCKey), and hierarchical support vector machines (LOCnet). Another dedicated nuclear sequence predictor, NucPred (Brameier *et al.*, 2007), has been recently developed to predict proteins that spend at least some time in the nucleus. NucPred is based on regular expression matching of NLSs and multiple program classifiers induced by genetic programming, and has similar overall prediction sensitivity and specificity with PSORT and PredictNLS. Predictors involving nuclear proteins also include NetNES (la Cour *et al.*, 2004)

that predicts nuclear export signal containing proteins.

While there are several dedicated tools that can directly predict or help identifying nuclear proteins, no particular prediction algorithm has been available that can predict proteins destined into the nucleolus or that can distinguish nucleolar proteins from nuclear proteins. Nevertheless, there has been studies to derive a knowledge-base that could be useful in predicting nucleolar proteins (Leung *et al.*, 2003), which generally suggested the use of amino acid and peptide composition and sequence homology information across different species.

4.1.1 Disordered protein regions

Natively unstructured regions are a common feature of eukaryotic proteins and many proteins have such regions with no well-defined 3-D structures in their native states (Dunker *et al.*, 2000). These natively unfolded protein regions could be involved in molecular recognition, and they can occasionally take regular forms when functioning. The first evidence came from a study carried out by Alber *et al.* in 1983, where it was concluded that the structure analysis of a complex, triose phosphate isomerase-substrate, had shown that a mobile region of 10 amino acids becomes ordered when an associated ligand binds. Disordered-to-ordered transition patterns can allow natively unstructured, related proteins to make formations (Weinreb *et al.*, 1996). However, these type of interactions involving disordered regions are not limited to only protein-protein interactions, and could be observed in protein-dna, enzyme-DNA, receptor-ligand interactions as well (Huber, 1979).

Dunker *et al.* (2000); Wright & Dyson (1999) showed that intrinsically unstructured protein regions are important regarding protein function. Lobley *et al.* (2007) directly used predicted disorder patterns successfully to improve protein function prediction. In Lokum, however, using disorder prediction did not improve the prediction of general localisation categories (see 3.3.7), so in this chapter, I try to address the potential contribution of protein disorder in distinguishing proteins localised in nucleoli from the rest of the other nuclear proteins, where I also used the features used in Lokum (Chapter 3).

4.1.2 Protein disorder region prediction

PONDR® is one of the best-known tools to predict disorder (Garner *et al.*, 1999; Li *et al.*, 1999; Radivojac *et al.*, 2003, 2004; Romero *et al.*, 2004). It uses pattern recognition techniques employing a set of attributes which are based on biological knowledge. Examples of other disorder software are FoldIndex (Prilusky *et al.*, 2005), DisEMBL (Linding *et al.*, 2003a), GlobPlot 2 (Linding *et al.*, 2003b), DISOPRED2 (Ward *et al.*, 2004), and Prelink (Coeytaux & Poupon, 2005).

The protein disorder prediction category has been introduced in the fifth “Critical assessment of methods of protein structure prediction” (CASP) competition (Cozzetto *et al.*, 2005, 2007; Soro & Tramontano, 2005; Valencia, 2005), with the participation of the mentioned programs and several others.

A program developed in 2005, RONN, has been recently compared with most of the notable CASP participants in the disorder category (Yang *et al.*, 2005) on an official CASP assessment dataset which contains 159 proteins sequences with experimentally determined disorder regions. Table 4.1 summarises the per-

formances of the 8 compared programs, with DisEMBL being compared using three different versions of the program. In addition to the traditional assessment measures sensitivity (Equation 2.3), specificity (Equation 2.4) and Matthew’s Correlation Coefficient (Equation 2.5), in CASP, a new weighted score (CASP-S) (Jin & Dunbrack, 2005) was used which was defined as:

$$CASP-S = \frac{100(w_{TP}TP + w_{FP}FP + w_{TN}TN + w_{FN}FN)}{TP + FP + TN + FN} \quad (4.1)$$

where w_{TP} stands for the number of disordered residues divided by the total number of residues, and so on. (w_{FN} was taken as $-w_{TP}$ and similarly, $w_{FP} = -w_{TN}$).

Also, the developers of RONN added yet another measure in their performance assessment, probability excess:

$$Prob. \ excess = \frac{TN \ TP - FN \ FP}{(FN + TP) + (TN + FP)} \quad (4.2)$$

Because of its reported reasonably good performance over the other predictors and availability as a stand-alone application I chose RONN for performing disordered protein region predictions.

4.2 Materials and methods

4.2.1 Datasets

The first proteins annotated as “nucleolar” came from mass-spectrometry studies (Andersen *et al.*, 2002, 2005; Scherl *et al.*, 2002). Recently, the list of nucleolar proteins, which have been previously identified mainly through mass-

spectrometry, has been extended by a protein-protein interactions approach (Hinsby *et al.*, 2006).

The nuclear and nucleolar protein sequences used in this study were downloaded from the LOCATE mouse protein sequence database (Fink *et al.*, 2006). LOCATE is a well curated, web-accessible database containing descriptions for the membrane organisation and subcellular localisation of FANTOM proteins. The FANTOM (Functional Annotation of the mouse) consortium (Carninci *et al.*, 2005; Maeda *et al.*, 2006) aims at providing the ultimate characterization of the mouse transcriptome. Only full length proteins from the FANTOM-3 project are present in LOCATE.

In LOCATE, I only considered protein annotations that are verified either by experiments or from literature. Among these, I picked nuclear (GO id:0005634¹)

¹<http://www.ebi.ac.uk/ego/GSearch?query=0005634&mode=id&ontology=component>

Method	SN	SP	MCC	Casp-S	Prob excess
RONN	0.603	0.878	0.395	9.33	0.481
DISOPRED2	0.405	0.972	0.470	7.81	0.377
PONDR®	0.557	0.816	0.278	7.22	0.373
DisEMBL(hot)	0.492	0.840	0.260	6.43	0.332
DisEMBL(465)	0.334	0.981	0.437	6.10	0.315
FoldIndex	0.488	0.811	0.224	5.79	0.299
PreLink	0.237	0.947	0.219	3.55	0.183
GlobProt	0.372	0.811	0.140	3.54	0.183
DisEMBL(coils)	0.740	0.424	0.104	3.19	0.165

Table 4.1: **Performance measures calculated from the blind testing of nine disorder prediction methods against the main blind test set of 80 proteins of CASP 6.** The performance measures are sensitivity (SN), specificity (SP), Matthews correlation coefficient (MCC), CASP S-score and probability excess (Prob. excess). This table is re-produced from Yang *et al.* (2005).

and nucleolar (GO id:0005730¹) protein sequences. I removed sequences annotated as both nuclear and cytoplasmic etc. to have two datasets at the end, one consisting of nucleolar proteins and another one consisting of only nuclear proteins. Some nucleolar proteins could also be annotated as nuclear, as they can spend some time in the nucleus, too. The final list of protein IDs used in this study can be found in Appendix C.

Using the CD-HIT (Li & Godzik, 2006) sequence clustering program to reduce the maximum sequence identity between any two sequences to 30%, the nuclear mouse protein dataset downloaded from LOCATE was reduced to 386 sequences from an initial number of 715. Similarly, the nucleolar set which initially had 815 sequences was filtered to allow a maximum identity of 30% between any two sequences at the end, which resulted in 397 sequences. One third of each dataset was reserved for testing purposes, while the remaining sequences were used in motif discovery.

Protein-capable NestedMICA (Doğruel *et al.*, 2008) was run on randomly chosen 257 nuclear and 265 nucleolar sequences, leaving the rest of the sequences in the datasets for test purposes. In order to detect possible motifs at both termini, N-terminal amino acid chunks of length 20 were compiled from the nucleolar and nuclear sequences. Similarly, two more datasets were produced which contained 20aa C-terminal sequences from both types. NestedMICA has been run on the nucleolar and nuclear training datasets containing whole-length sequences, 20 N-terminal amino acid chunks, and finally 20 C-terminal peptides.

¹<http://www.ebi.ac.uk/ego/GSearch?query=0005730&mode=id&ontology=component>

4.2.2 Training background models for nucleolar and nuclear datasets

Two NestedMICA background models were trained using a similar strategy described in the background model related sections on pages 35 and 41. However, particularly the nuclear motifs (shown in Figure 4.2) obtained using these background models were quite short and surprisingly not rich in residues like Lysine or Arginine which are expected to be abundant in the core parts of the NLSs. Nucleolar motifs were quite short, too, and they only possessed strong Arginine residues but no Lysines (Figure 4.2). This could have resulted because of using relatively simple, zero order background models which are trained on the relatively small number of sequences in these two datasets (in Chapter 2 I showed that using order-1 background models would be better than using an order-0 background, but if there is enough data to train it).

An alternative, third background model was trained using 438 redundancy reduced cytoplasmic sequences (see page 55). Nuclear proteins are transferred into the nucleus by the means of some molecules binding to their NLSs. Therefore, as previous studies have shown, for example by Goldfarb *et al.* (1986), if these signals are altered it is likely that a protein will remain in the cytoplasm and will not be able to be carried into the nucleus. Thus, the uninteresting, non-localisation segments of nuclear and other sub-nuclear proteins could best be represented by a cytoplasmic background. Indeed, when I ran NestedMICA with this first order cytoplasmic background model consisting of 4 mosaic classes on the individual nuclear and nucleolar protein datasets that have been created, the results were much more promising. Motifs obtained from each background model







Localisation	Sequence segment	Motif
a) Nucleolar	N-terminal	
b) Nuclear	N-terminal	
c) Nucleolar	Entire	
d) Nuclear	Entire	
e) Nucleolar	C-terminal	
f) Nuclear	C-terminal	

Figure 4.2: **NestedMICA motifs discovered from nuclear and nucleolar datasets.** NestedMICA was run on two sets: nuclear and nucleolar datasets. In each run, it used a dedicated background model trained with the corresponding dataset. Figure 4.4 shows a set of “better motifs” discovered using another (cytoplasmic) background model trained with more sequences.

were assessed in terms of their performances to separate nuclear and cytoplasmic proteins, and it was actually when I used this cytoplasmic background model that they better discriminated the two classes, rather than when I tried the two background models trained on nuclear and nucleolar sequences.

4.2.3 Running RONN

The RONN protein disordered prediction program (Yang *et al.*, 2005) was run with the “short output” command line options on a Linux server. BioJava scripts were written to parse the output of the program and perform the statistics. A score of greater than 0.5 was considered as a disordered prediction, as recommended in the RONN manual. Figure 4.3 shows an example plot drawn according to RONN predictions from a nuclear sequence, where RONN produces disorder scores for each amino acid position. RONN version 3 was obtained by personal communication with the program’s developers.

4.2.4 Training the SVM

As in previous chapters, a popular Support Vector Machine (SVM) implementation, *libsvm* (Chang & Lin, 2001), was used in the task of classifying nuclear and nucleolar proteins. 10-fold cross validation was applied, and I used a radial-basis kernel function whose gamma (γ) and C penalty parameters were systematically optimised.

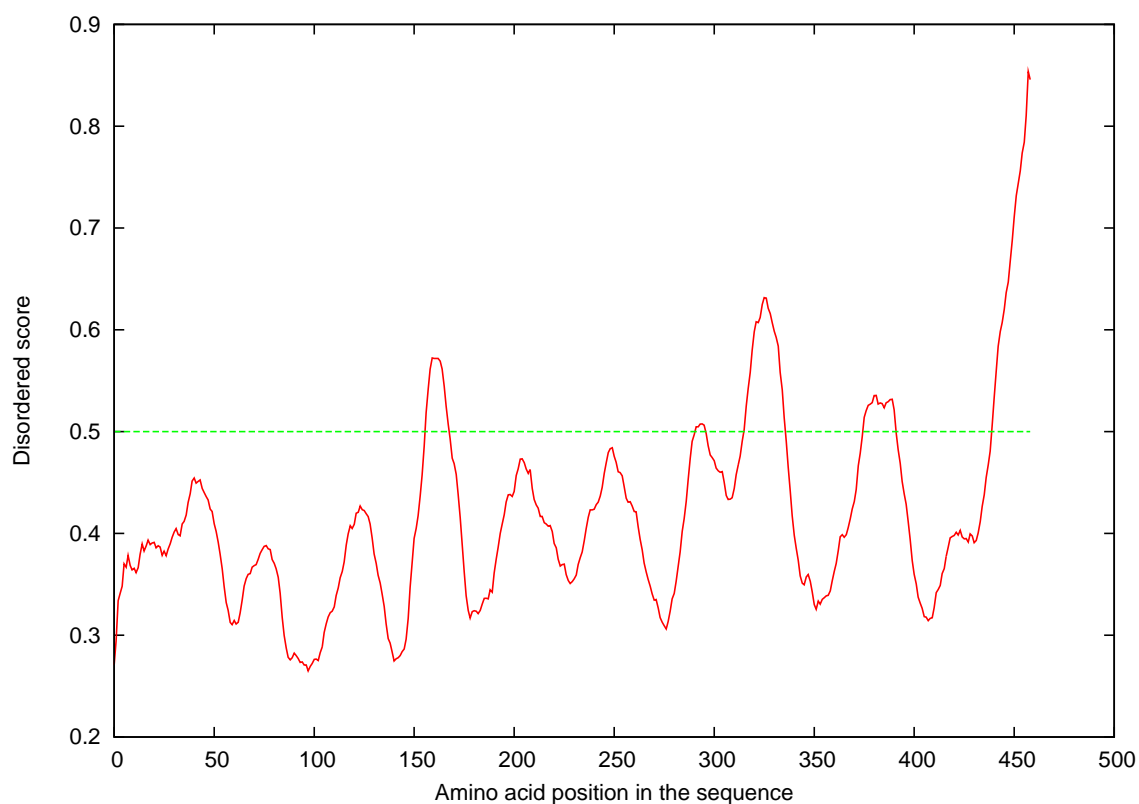


Figure 4.3: **A protein disordered region plot based on RONN predictions.** The plot shows the disorder score of a nuclear sequence of length 459, as an example. RONN produces a score between 0 and 1 for every single amino acid position across a sequence. A score above 0.5 indicates a disordered residue or a region. As the plot illustrates, a sequence can have multiple disordered regions (5 in this example, with a strong disordered sequence chunk at the C-terminal end).

4.3 Results

Nucleolar proteins possess NOSs (Dang & Lee, 1989) to enter into the nucleus from the cytoplasm. Figure 4.4 shows some of the nuclear and nucleolar protein motifs reported by NestedMICA. NestedMICA was run on 3 datasets for each localisation class: a dataset consisting of full-length sequences, and two datasets of 20aa N- and C-terminal sequence chunks, respectively. The most striking difference between the nuclear and nucleolar sequence motifs is how nucleolar motifs are enriched with Arginine (R) and Lysine (K) amino acid letters over the nuclear motifs discovered in the N- and C-terminal regions. NLSs have been known not to have specific positions and can be located across the entire primary structures of nuclear proteins; however, these results suggest the possibility that nucleolar proteins, unlike nuclear proteins, have stronger NLS-like motifs (NOSs) in their both N and C termini. We scanned and scored both the N- and C-terminal nucleolar motifs (Figure 4.4) in the corresponding 20 aa N or C terminal regions of both nucleolar and nuclear proteins to see if we can observe any difference in the score distributions. The highest scores obtained from these nucleolar motifs both in the nuclear and nucleolar sequences are plotted in Figure 4.5.

By using a simple SVM consisting of input vectors formed with only the scores of the N- and C-terminal motifs (4 in total) shown in Figure 4.4, it was possible to classify 65.4% of the proteins correctly into the two classes of nuclear and nucleolar localisations. Adding amino acid composition to the four motif scores, I was able to increase the performance up to 74.5%. Using only the 20-dimensional amino acid composition rates was sufficient to predict 73.5% proteins correctly.







Localisation	Sequence segment	Motif
a) Nucleolar	N-terminal	
b) Nuclear	N-terminal	
c) Nucleolar	Entire	
d) Nuclear	Entire	
e) Nucleolar	C-terminal	
f) Nuclear	C-terminal	

Figure 4.4: A selection of the protein motifs recovered by NestedMICA from a set of nuclear and a set of nucleolar proteins, using a cytoplasmic background. N-terminal motifs shown were reported from the first 20 N-terminal amino acid regions. Similarly, the C-terminal motifs were searched within the last 20 amino acid regions. Other motifs indicated by the “Entire” segment were discovered when full length sequences were used.

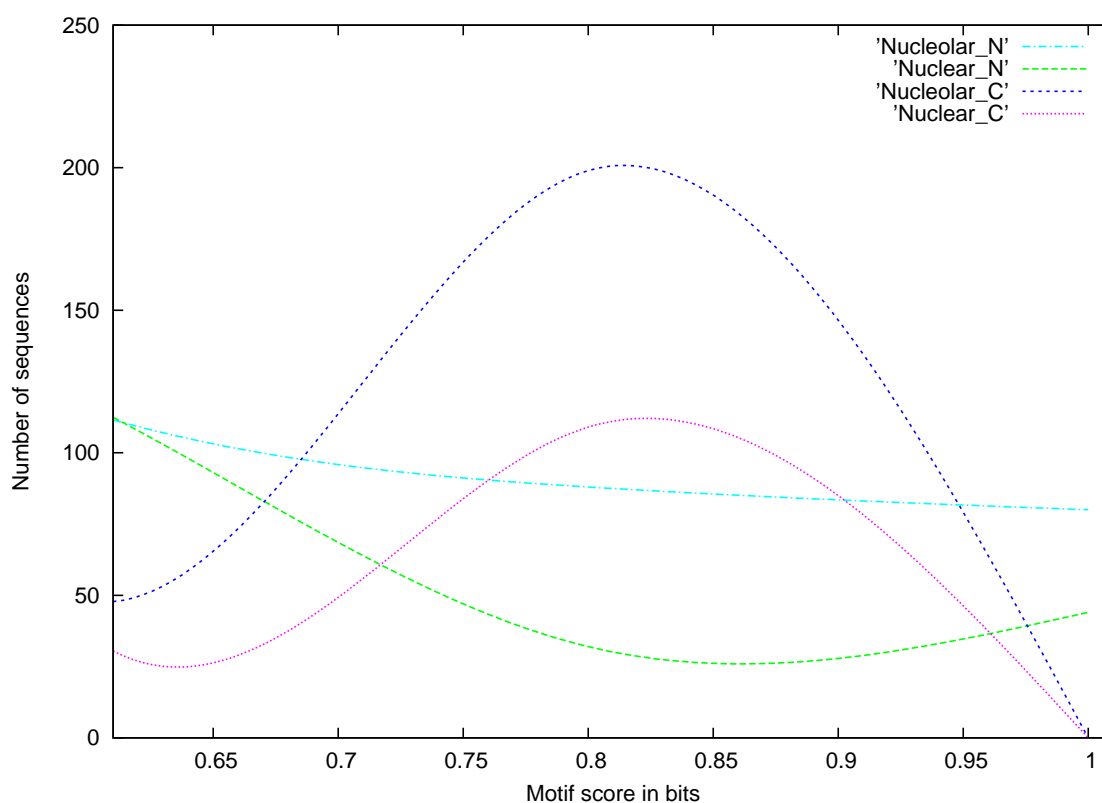


Figure 4.5: **Score histograms for N- and C-terminal nucleolar motifs.** Both nuclear and nucleolar sequences were scanned using the N-terminal nucleolar motif (Figure 4.4a). Similarly, the C-terminal nucleolar motif (Figure 4.4e) was scored in both types of datasets. Scores shown on the x-axis correspond to the best matches within the relevant 20 amino acid long N or C terminal chunks. The C-terminal motif generated Gaussian-like distributions when scored in the last 20aa C-terminal regions, however, this motif is clearly more abundant in the nucleolar C-termini. The other two curves indicate that the N-terminal regions are less abundant in terms of the N-terminal nucleolar motif, but still, this motif was less frequent in the N-termini of nuclear proteins than the N-termini of proteins localised in the nucleoli.

Using transmembrane (TM) statistics reported by TMHMM (Krogh *et al.*, 2001) (reported “features” are summarised in 3.2.5) improved the prediction accuracy in Lokum. Furthermore, nuclear proteins, in theory, should possess a larger number of TM helices, compared to the nucleolar sequences which are confined to the centre of the nucleus and less likely to have TM helices. In fact, running TMHMM on the entire sequences in both datasets to compare them in terms of their number of predicted residues that possibly lie in a TM helix (Figure 4.6) revealed that this feature can significantly improve predictions. With the addition of the two more types of predicted TM statistics mentioned in Section 3.2.5, the correct prediction rate increased to 77.14%. When I used the three TMHMM statistics alone, the correct prediction rate was 64.4%.

Finally, after adding the bipartite NLS motif (Figure 3.10, page 92) that we obtained using the combinatorial approach involving both NestedMICA and Eponine, the overall correct classification rate increased to as high as 78.42%. Sequences used in the motif discovery were not used in training and testing of the SVM. Due to the relatively low number of sequences in both datasets (783 in total), this particular SVM was trained and tested using 10-fold cross validation (see Methods).

That amino acid composition helped us in making more correct predictions implies there is a certain degree of bias in composition even between the similar classes of nuclear and nucleolar proteins, which could be associated with the possibility that nucleolar proteins have slightly different compositional preferences than the other proteins in the nucleus so as to allow them to be packed more tightly to form the nucleolus. Figure 4.7 shows the compositional differences

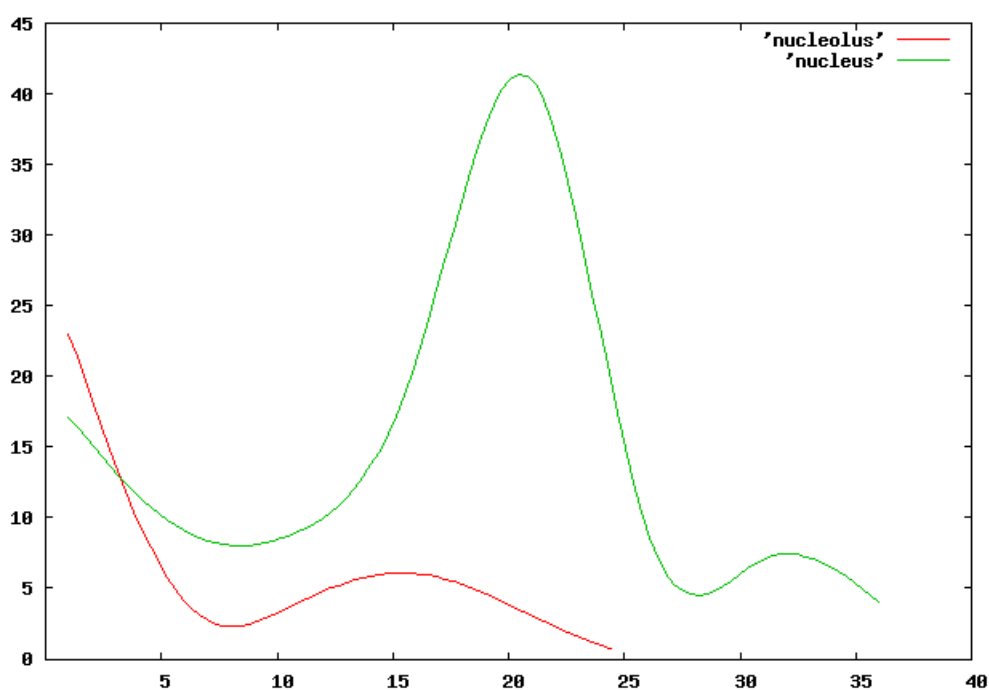


Figure 4.6: **Distributions of amino acids predicted to be within TM helices in nuclear and nucleolar proteins.** TMHMM (Krogh *et al.*, 2001) was run on the entire nucleolar and nuclear protein sequences. The curves show the total number of sequences (y-axis) having a certain, predicted total number of amino acid residues in their sequences that fall in a membrane-spanning region, for nucleolar (red), and nuclear (green) proteins. According to this plot, most nuclear proteins have around 20 amino acids within their TM helices all together. A bin size of 5 amino acids was used to plot the frequencies, and the curves were smoothed by the “cubic splines” algorithm.

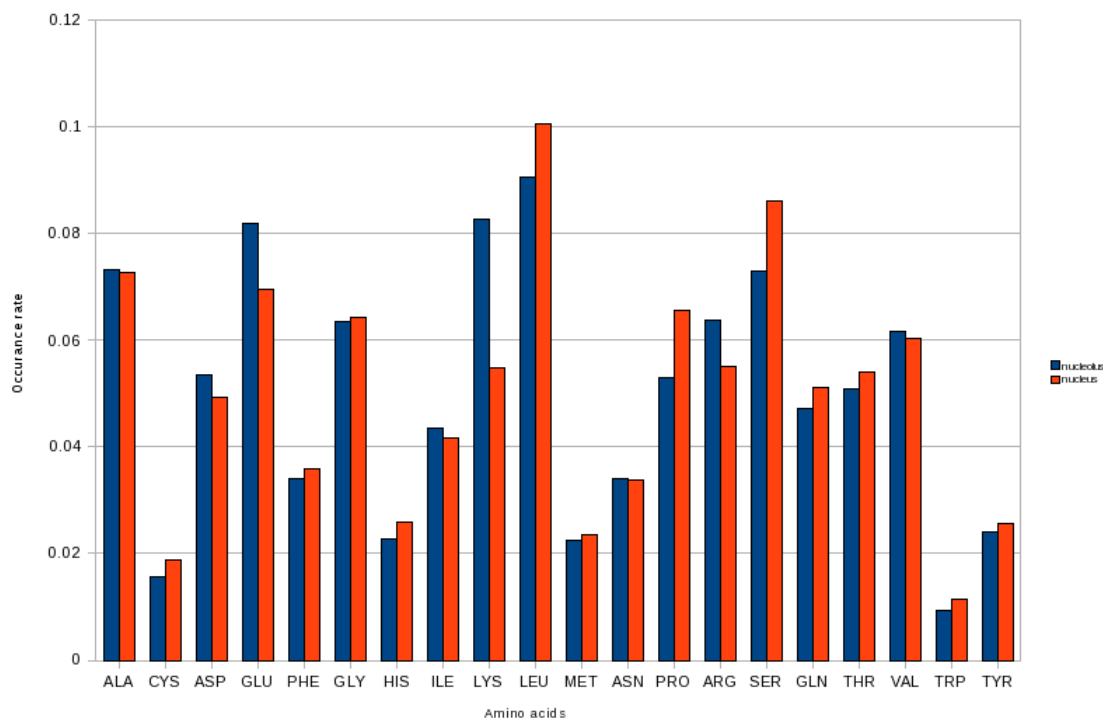


Figure 4.7: **Differences between nucleolar and nuclear proteins in terms of their amino acid compositions.** This statistics was obtained using the nuclear sequences datasets consisting of 386 sequences and the nucleolar sequence set having 397 sequences (see Materials and Methods). The most noticeable difference is how nucleolar proteins are enriched with Lysine (K) over nuclear proteins.

between the two types of proteins localised in nuclei and nucleoli. A similar figure showing the comparison of nucleolar and nuclear proteins in terms of their amino acid composition has been reported previously by [Leung *et al.* \(2003\)](#). As seen in Figure 4.7, the most notable difference is how nucleolar proteins are enriched with Lysine (K) over nuclear proteins. While most other amino acid composition rates were more or less identical, nuclear proteins had a larger number of the nonpolar amino acids Leucine (L) and Proline (P), and the polar Serine (S) than the nucleolar proteins.

The fact that our motif finder discovered some motifs (Figure 4.4) from the nucleolar protein set does not necessarily mean that these motifs can not be found in the nuclear proteins, and vice versa. As can be seen in Figure 4.8, which, as an example, shows the score distributions of motif c of Figure 4.4 for both types of protein sequences, some nuclear proteins may also contain this particular K- and R-rich motif despite that it was originally discovered in the nucleolar set. However, the histogram plot suggests that mostly high scoring instances of this motif are more abundant in nucleolar proteins compared to the best hits of the motif in sequences localised in the nucleus.

In addition to demonstrating that terminal regions of nucleolar proteins could be more biased towards positively charged residues, I investigated whether nuclear and nucleolar proteins differ in terms of their disordered region distribution. 41.99% of the amino acids in the nucleolar proteins set and 41.87% of the amino acids in the nuclear proteins set were predicted as disordered by the RONN software. This indicates that there is no significant difference in terms of the number of residues falling into a disordered region between both types of proteins. However, there is a difference about what constitutes these disordered regions: it turned out that these disordered regions are enriched more with charged residues in proteins localised in the nucleolus over proteins of the nucleus after some tests I performed with some of the motifs found.

Using motif c of Figure 4.4 to scan only both types of sequences, I observed that a larger number of disordered regions in the nucleolar sequences contained this motif than disordered regions in the nuclear proteins. Figure 4.9 shows the normalised frequency distribution of strong hits of this charged-residue rich

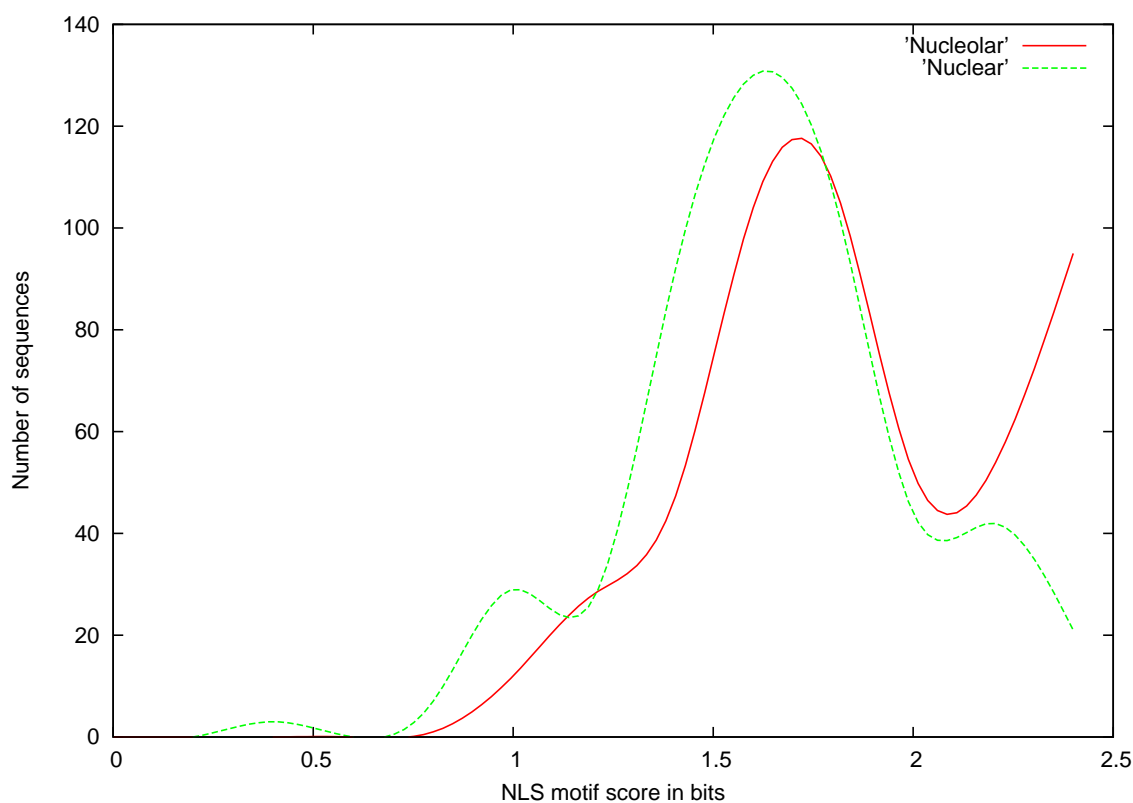


Figure 4.8: **Score distribution of a core nucleolar motif within nuclear and nucleolar proteins.** Motif scores shown on the x-axis are given in information bits, for the best match per sequence. The y-axis indicates the number of sequences for nucleolar (red line) and nuclear (green dashed line) featuring this motif with different scores. For plotting the histogram, 500 nucleolar and 500 nuclear sequences that were sampled randomly from the original datasets have been used.

motif within the predicted disordered regions for both types of sequence classes. Sequence regions scoring less than an empirically chosen value of 1.8 were not considered as true NLS matches (Figure 4.8). However, a second similar analysis performed by using another motif, which was discovered from a general nuclear localisation dataset in the previous chapter (Figure 3.4i), revealed that even when we consider the entire range of scores without using any threshold it is still possible to observe the same kind of tendency of finding more NLS motifs within disorder regions (Figure 4.10).

Given that there is a tendency in nucleolar proteins to possess “K & R”-rich motifs more abundantly within their disordered regions compared to nuclear proteins, I investigated whether this bias could be used in a prediction system. The SVM that was built initially to distinguish nuclear proteins from nucleolar proteins was modified so as to allow us to test this phenomenon. To this end, firstly, I added to the SVM the best scores of those core NLS signals (represented as a PWM in part c of Figure 4.4) that fall into a disordered region, excluding other potential motif hits in the rest of the sequence regions. Secondly, I added the predicted disordered scores of sequence regions featuring a core nucleolar motif, such as motif c of Figure 4.4. Unfortunately, both approaches, when used separately or together, failed to provide a substantial increase in the SVM’s performance, meaning that their potential contributions are somehow already achieved by the other used features including NLS motif scores and amino acid composition etc. Using general disordered statistics for each sequence, such as the number of disordered blocks per residue and the ratio of amino acid residues predicted as disordered to the total number of residues in a sequence, resulted

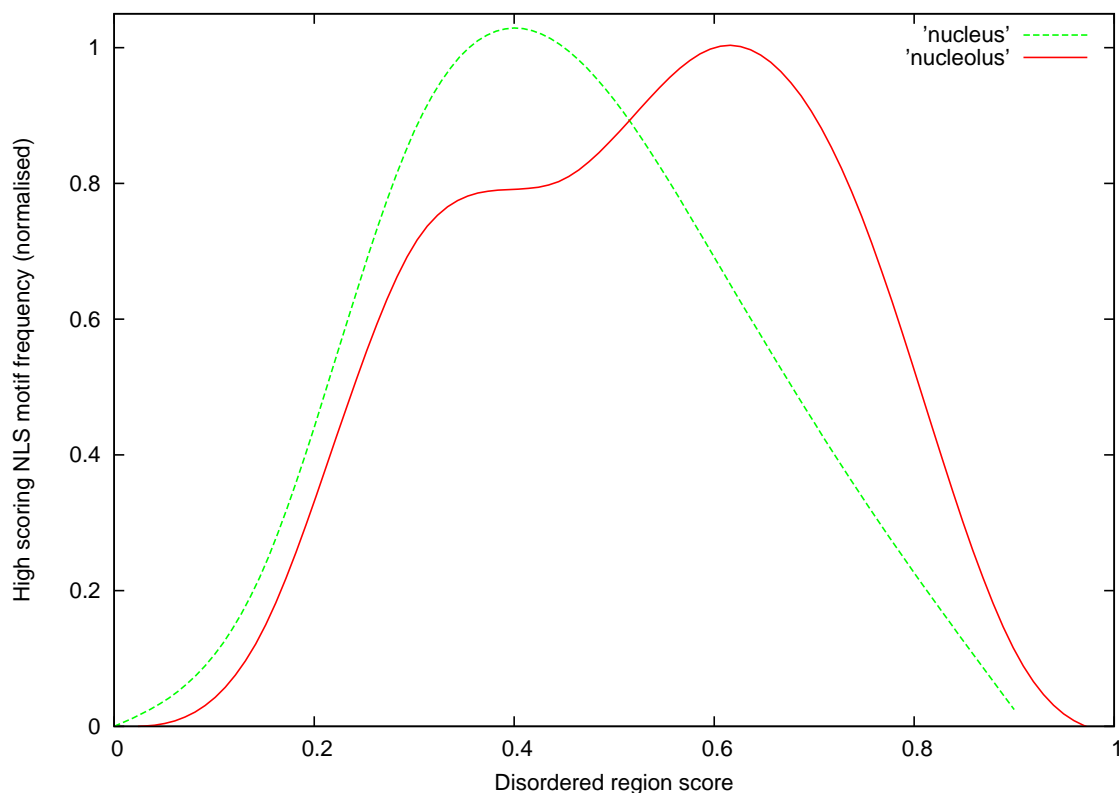


Figure 4.9: **A larger number of nucleolar localisation motif hits fall in disordered regions, compared to the NLS motifs in disordered regions of nuclear proteins.** The y-axis corresponds to the normalised frequencies, while the x-axis represents the disorder region scores as reported by RONN. A score of greater than 0.5 indicates a predicted disorder region. The dashed green curve represents nuclear proteins which show a normal distribution around a score of 0.4, while the solid red curve shows the histogram for nucleolar sequences having a tendency to contain more number of the NOS within their disordered regions.

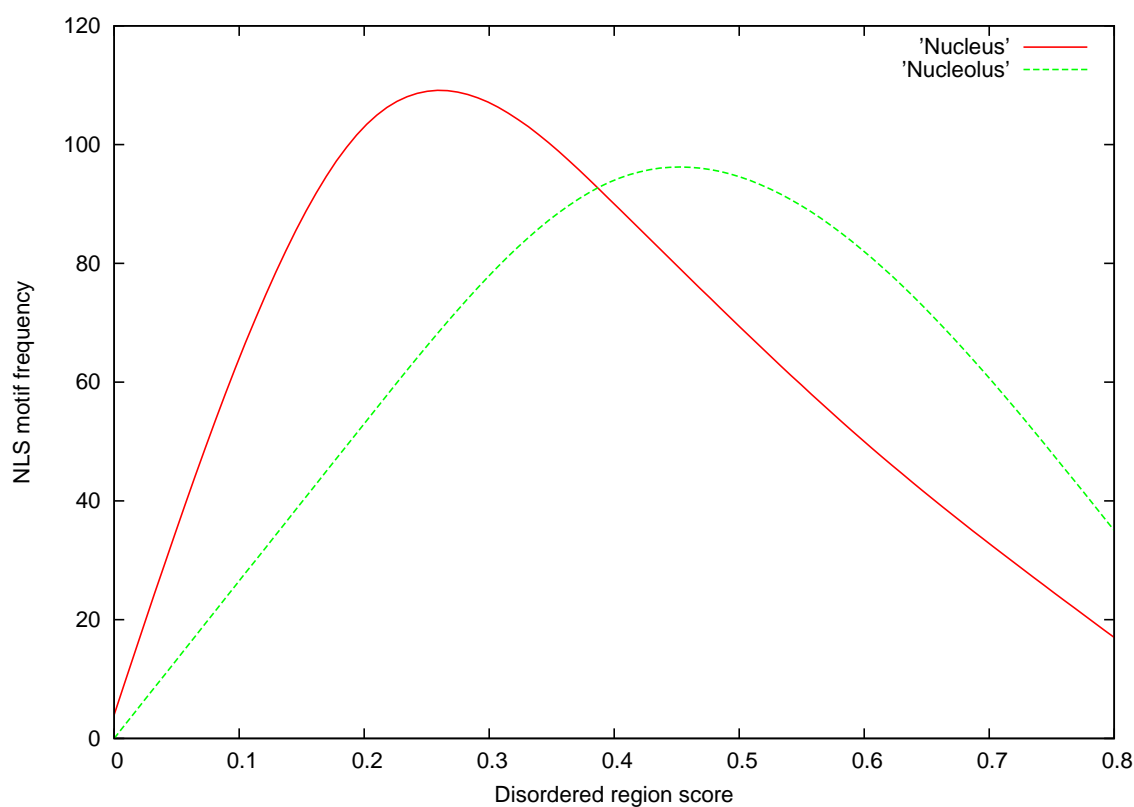


Figure 4.10: **Generally a larger number of NLS signal hits can be found in disordered regions of nucleolar proteins compared to nuclear sequences.** The y-axis corresponds to the frequencies of motif hits, while the x-axis represents the disordered score regions as reported by RONN. A score greater than 0.5 indicates a predicted disordered region.

in the same maximum correct prediction percentage (78.42%) that I obtained without using the disorder-related scores (see above). At the end, scores associated with disorder were not included in the SVM, as this did not improve the performance, although the nucleolar sequences showed a bias to possess a larger number of “K & R”-rich motifs in their disordered regions.

4.4 Discussions and conclusions

Using the observation that nucleolar proteins tend to contain a larger number of charged residues in their disordered regions was not particularly helpful in automatic classification of nuclear and nucleolar proteins. Instead, using these motifs directly without considering disordered regions to score proteins was more effective. In addition to using the reported motifs found in the terminal regions of nucleolar and nuclear sequences, incorporating amino acid composition in the SVM proved useful, as in predicting major localisation categories (see previous chapter). Thus, despite being confined by the nuclear membrane and sharing similar characteristics, there are significant differences in amino acid compositions between the members of these two types of proteins.

Loop regions and regions with no specific secondary structure in proteins do not have to be disordered necessarily. A disordered region means that that region has the capacity to change into an ordered state when needed, unlike, for example, some loop regions which can not become “ordered”, that is, have a certain structure and shape.

It is not very surprising to have observed that proteins forming the subnuclear compartment nucleolus are rich in charged amino acids like K and R, and that

they are more abundant in regions predicted to be disordered. It has been shown that aromatic amino acids like Tryptophan (W), Tyrosine (Y) and Phenylalanine (F) are less likely to be found in long disordered regions (Kissinger *et al.*, 1995), because these amino acids usually have a strong interaction capability to develop a structure, and thereby they inhibit disorder (Burley & Petsko, 1985). It has also been observed by the same groups that charge imbalance in protein sequences tends to favour disorder. But to find out that there are more of these charged residues in disordered regions of nucleolar proteins compared to nuclear proteins was surprising. This can be explained, to a certain extent, by the speculation that nucleolar proteins have to behave like any other nuclear proteins while traversing the nuclear pore to enter into the nucleus, but after that point, most probably their disordered regions which potentially convey the extra signals of nucleolar localisation signals (NOS) involved in their transport into their subnuclear destination, become more ordered and functional.

Unfortunately, good quality and reliable localisation annotation is too limited to satisfactorily study sub-localisation classes such as nucleolar or mitochondrial membrane proteins. Also, there can always be annotation errors in the datasets used. I tried to minimise these data related problems by choosing manually annotated and well curated datasets. To avoid a potential bias in predictions, sequence identity was lowered to a maximum of 30% by using a clustering algorithm (see methods). Another problem stemming from the underlying biology is that some proteins can be functioning in more than one compartment. However, even if the datasets contain such protein sequences, statistically it should still be possible to retrieve the general characteristics representing an individual group.

In the case of motif finding, for example, a few sequences coming from different types of protein localisations or those having multiple possible localisations should not prevent NestedMICA from finding the overexpressed, representative sequence motifs.

In spite of possible errors and data related limitations, I think the observation that disorder regions have more charged residues in nucleolar proteins, the compositional differences between the two classes, and finally the motifs found in the terminal sequence regions to distinguish nuclear and nucleolar proteins are promising results that can be used for discriminating nucleolar proteins from other nuclear proteins, as the tests indicated.