ARC Centre for Complex Systems, Australia

www.accs.edu.au

Modelling the Import of Nuclear Proteins

ARC CENTRE FOR COMPLEX SYSTEMS

John Hawkins & Mikael Bodén

ARC Centre for Complex Systems
School of ITEE | The University of Queensland | ST LUCIA QLD 4069 | AUSTRALIA

T: +61 7 3365 1003 | F: +61 7 3365 1533 | E: outreach@accs.edu.au

Introduction

- After being manufactured in the cytosol, many proteins undergo transport to organelles.
- The transport processes rely on signals within the protein.
- The signals are recognised by organelle specific mechanisms, which then effect transport.
- Ideally we would like to be able to predict the localisation of a protein from primary structure alone.

ARC CENTRE FOR
COMPLEX SYSTEMS

Nuclear Import and the Regulatory Proteome

- Other organelles tend to have stable protein content
- The nucleus has a far more transitory protein complement
  - Due to the fact that localisation is one of the mechanisms of regulating interaction with the genome
- Hence accurate nuclear localisation prediction must form part of our understanding of the regulatory proteome.

ARC CENTRE FOR
COMPLEX SYSTEMS

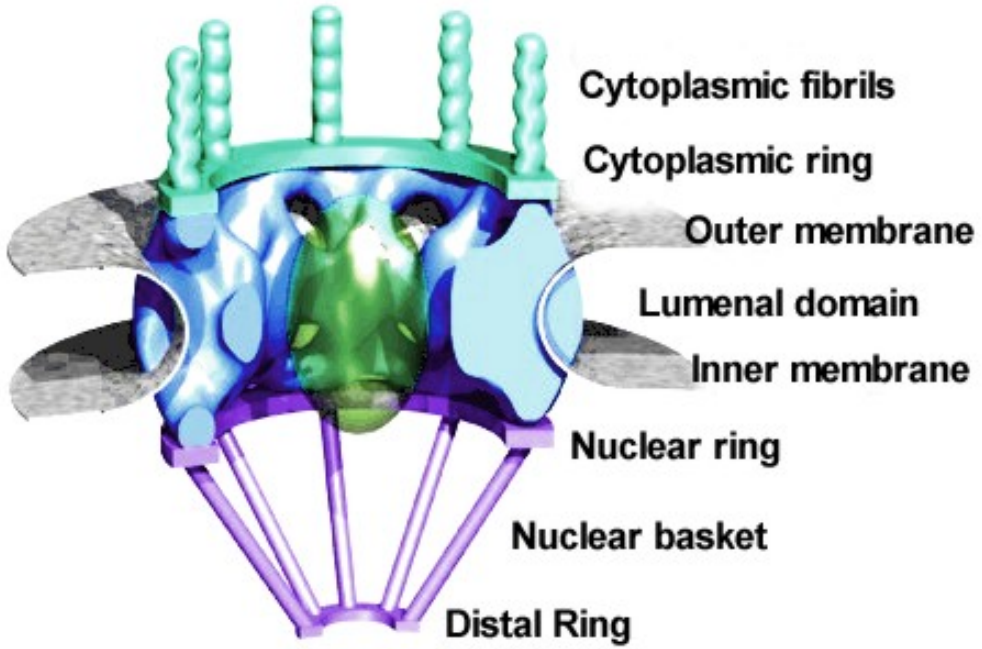
Nuclear Localisation

- Nuclear Localisation is unusual
  - Unlike some other localisation processes does not use a cleaved targeting peptide.
  - Large number of signals with no defined position.
  - Proteins are imported in their mature folded state.
- Transport requires a large channel through the nuclear membrane.
  - A macromolecule called the Nuclear Pore Complex

ARC CENTRE FOR
COMPLEX SYSTEMS

Nuclear Pore Complex (NPC)

- Spans the membrane between nucleus and cytoplasm.
- Passive diffusion of small molecules
- Large molecules require assistance.



ARC CENTRE FOR
COMPLEX SYSTEMS

Getting Imported

- Mediated import through the NPC requires the assistance of proteins called Importins.
- These recognise the Nuclear Localisation Signal (NLS) on the ‘cargo’ molecule and bind to it.
- The importin-cargo complex binds to the cytoplasmic fibrils.
- Conformational change in the fibrils effects the transport of the importin-cargo complex.

Classical Nuclear Localisation Signals

- Usually short stretches of basic amino acids.
- SV40 large T-antigen like NLSs:
  - Pro-Lys-Lys-Lys-Arg-Lys-Val
- Xenopus nucleoplasmin like NLSs:
  - Two basic regions separated by a spacer.
- Mat α2 amino terminal signal:
  - Lys-Ile-Pro-Ile-Lys

ARC CENTRE FOR
COMPLEX SYSTEMS

Ambiguities

- Matching to the ‘classical’ NLS is very flexible.
- Likely to occur reasonably frequently
- Hence not a reliable prediction method.
- In recent years numerous other highly specialised NLSs have been discovered.

ARC CENTRE FOR
COMPLEX SYSTEMS

Non-Classical NLSs

- Christophe *et al* (2000) Give a listing of non-classical sequences known to be involved in nuclear localisation.
  - varying lengths up to 38 residues.
  - None of which are particularly high in basic residues.
  - One example they site is high in acidic residues.
- NLSdb – Nair, Carter & Rost (2003)
  - Contains 308 entries.
    - 114 Experimentally determined NLSs
    - 194 Discovered through *in-silico* mutagenesis

ARC CENTRE FOR
COMPLEX SYSTEMS

Machine Learning Models

- Models that predict Nuclear Localisation rely on information other than the presence of these known NLSs.
- Some form of sequence summary
  - K-mer
    - Amino Acid composition
    - Di-peptide composition
- In recent years a range of sophisticated ensemble models have emerged that include the nucleus among their predicted localisations.

ARC CENTRE FOR
COMPLEX SYSTEMS

Independent Test Results

Model	Dataset	Accuracy	Sensitivity	Specificity	MCC
Nucleo V1.0	Dual	0.64	0.68	0.62	0.28
	Non-Dual	0.71	0.80	0.62	0.42
	All	0.70	0.76	0.62	0.38
LOCtree	Dual	0.64	0.55	0.68	0.22
	Non-Dual	0.66	0.64	0.68	0.32
	All	0.64	0.61	0.68	0.29
P2SL	Dual	0.62	0.35	0.74	0.10
	Non-Dual	0.62	0.49	0.74	0.24
	All	0.57	0.45	0.74	0.19
HSLPred	Dual	0.67	0.56	0.71	0.25
	Non-Dual	0.57	0.43	0.71	0.14
	All	0.57	0.46	0.71	0.18
predictNLS	Dual	0.71	0.21	0.93	0.20
	Non-Dual	0.62	0.30	0.93	0.29
	All	0.54	0.27	0.93	0.25
SubLoc	Dual	0.57	0.66	0.53	0.17
	Non-Dual	0.64	0.75	0.53	0.29
	All	0.64	0.72	0.53	0.25

Problems

- Incomplete domain knowledge
- Existing predictors often use training sets consisting of exclusively nuclear localised proteins.
  - However 20% of known nuclear proteins are dual localised.
- Some models rely on protein homology
  - Hence have questionable capacity to generalize to truly novel sequences

ARC CENTRE FOR
COMPLEX SYSTEMS

An Ideal Model

- An ideal model for Nuclear prediction would be able to base its predictions on the correct identification of the NLSs
- To do this the model needs to be able to
  - Recognise the motifs in the sequence
  - Regardless of position
  - Recognise positional context
  - Allow for dependencies between regions

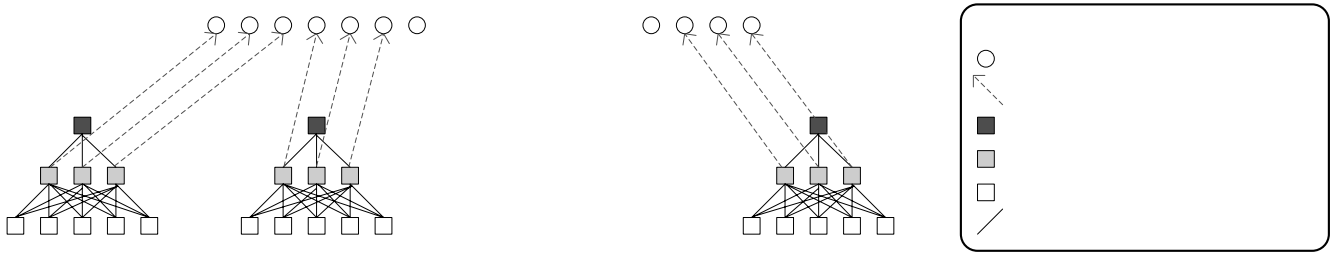
ARC CENTRE FOR
COMPLEX SYSTEMS

The PALFE Framework

- Proportionally Assigned Local Feature Extraction framework of neural networks
- Using an array of neural networks
- Each assigned a ‘local’ window of the sequence
  - Based on a proportion of the sequence length
- Each networks is trained to attempt the global task
  - Thus learning the ‘local’ features most pertinent to the ‘global’ task

The PALFE Framework cntd

- Each ‘local’ feature extraction network generates a feature vector in the hidden node activations of the network
- The feature vectors are concatenated to produce a global feature vector for the sequence



ARC CENTRE FOR
COMPLEX SYSTEMS

Simulation Protocol

- Use feature vector to train SVM
- Five fold cross validation
- Compared against:
  - A set of standard sequence kernels
  - A normalised sequence fed into an SVM
  - A null model consisting of the concatenated input windows fed into an SVM

ARC CENTRE FOR
COMPLEX SYSTEMS

Cross Validation Results

Model	Configuration	Accuracy	Sensitivity	Specificity	MCC
Mismatch	k=3,m=1	0.675	0.662	0.689	0.351
Substitution	k=3	0.686	0.671	0.703	0.374
Local Alignment		0.676	0.715	0.634	0.35
Spectrum	k=4	0.711	0.669	0.751	0.427
Spectrum	k=5	0.704	0.642	0.753	0.415
Composite Spectrum	k=1,4,5	0.749	0.759	0.76	0.497
Normalised	L=10	0.717	0.775	0.709	0.433
NULL-PALFE	10*31	0.658	0.72	0.657	0.313
PALFE FFNN	10*31	0.673	0.701	0.681	0.343
PALFE BiD-RNN	10*31	0.824	0.837	0.828	0.648

ARC CENTRE FOR
COMPLEX SYSTEMS

PALFE Performance

- The BiD-RNN PALFE networks provide superior performance to all other models
- The FFNN PALFE networks perform only marginally better than the NULL model
  - And worse than the normalised sequence
- We have conducted several analyses of the structure of the BiD-RNN feature space
  - Suggests that they are better equipped to accommodate movement in the position of the motifs

ARC CENTRE FOR
COMPLEX SYSTEMS

Independent Test Again

Model	Dataset	Accuracy	Sensitivity	Specificity	MCC
Nucleo V2.0	Dual	0.67	0.86	0.59	0.42
	Non-Dual	0.78	0.97	0.59	0.60
	All	0.79	0.93	0.59	**0.57**
Nucleo V1.0	Dual	0.64	0.68	0.62	0.28
	Non-Dual	0.71	0.80	0.62	0.42
	All	0.70	0.76	0.62	0.38
LOCtree	Dual	0.64	0.55	0.68	0.22
	Non-Dual	0.66	0.64	0.68	0.32
	All	0.64	0.61	0.68	0.29
P2SL	Dual	0.62	0.35	0.74	0.10
	Non-Dual	0.62	0.49	0.74	0.24
	All	0.57	0.45	0.74	0.19
HSLPred	Dual	0.67	0.56	0.71	0.25
	Non-Dual	0.57	0.43	0.71	0.14
	All	0.57	0.46	0.71	0.18
predictNLS	Dual	0.71	0.21	0.93	0.20
	Non-Dual	0.62	0.30	0.93	0.29
	All	0.54	0.27	0.93	0.25

ARC CENTRE FOR
COMPLEX SYSTEMS

Conclusion

- Our Final Model,
  - Accuracy of 0.79
  - MCC of 0.57
- According to our independent test it is the best current model of nuclear import prediction.
- We are now approaching the task of distinguishing between the dual and statically localised nuclear proteins.
- This will allow use to begin to focus in on predicting the regulatory proteome

ARC CENTRE FOR
COMPLEX SYSTEMS