Quantification of subcellar localization and co-localization from high-content images

Zhu, Shiwen

2015

Zhu, S. (2015). Quantification of subcellar localization and co-localization from high-content images. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/65583>

<https://doi.org/10.32657/10356/65583>

Quantification of Subcellular Localization and Co-localization from High-content Images

SHIWEN ZHU
(B.S., PEKING UNIVERSITY)

A THESIS SUBMITTED
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN COMPUTATION AND SYSTEMS BIOLOGY (CSB)
SINGAPORE-MIT ALLIANCE
NANYANG TECHNOLOGICAL UNIVERSITY
2015

Acknowledgment

I never expected it would take me that long to reach here. At this moment, I would first like to thank my parents for their endless love and support for me to overcome all the difficulties during my candidature. This thesis is dedicated to them.

I would like to express my special thanks to my supervisor, Paul Matsudaira for his guidance, encouragement, and inspiration during the past years. I am grateful for his continuous support and patience when I encountered difficulties and struggled to find my research direction.

I am grateful to my supervisor, Prof. Roy Welsch for his guidance and support, although there is a long distance from Singapore to MIT. I have surely inspired a lot from his advice and guidance.

I would like to express my gratitude to my supervisor Prof. Kwoh Chee Keong. His support and encouragement was quite important to me at the crucial period of my candidature.

I wish to express my sincere thanks to Prof. Jagath Rajapakse, who has given me abundant advice and support during my candidature, especially the period when I was trying to define my thesis topic.

I appreciate the opportunity provided by Singapore-MIT Alliance to pursue my Ph.D study at Singapore and Boston and I am grateful for the administrative support during the past years.

I joined SMA programme with a group of enthusiastic colleagues, Shuoyu, Junjie, Li Yan, Yin Lu, Shi Yuan, and we became best friends in these years. I really enjoyed my moments with them and am grateful to their encouragement and help.

Last but not least, it is always a pleasure to work with these wonderful people from different labs, Fang Fang, Aikia Yip, Tong Yan, Yi-Jen Chiu, Nikhil Jain, Dipanjan, Siew Ping Han, Sahar, Nicolas Boichat, Zhong Jun, Keshma from CBIS NUS, Piyush, Merlin, Liu Song, Jierong Cheng, Alvin from BIRC NTU, and James Evans from MIT. I would like to express my appreciation to all of them.

Publications

Papers

Zhu, S., P. Matsudaira, R. Welsch and J. C. Rajapakse (2010). Quantification of cytoskeletal protein localization from high-content images. Pattern Recognition in Bioinformatics, Springer: 289-300.

S. Zhu, C.K. Kwoh, J. C. Rajapakse, R. Welsch, P. Matsudaira “A Computation Method for Cytoskeleton Protein Localization Prediction from High-Content Microscopic Images”, (In preparation)

Presentations

S. Zhu, P. Matsudaira, P. T. C. So, J. C. Rajapakse “Analyzing Cytoskeletal Protein Colocalization from High-Content Images”, SMA symposium 2010 (Poster presentation).

S. Zhu, P. Matsudaira, R. Welsch, J. C. Rajapakse “Quantification of Cytoskeletal Protein Localization from High-Content Images”, 5th IAPR International Conference on Pattern Recognition in Bioinformatics (PRIB 2010), The Netherlands (Oral presentation).

S. Zhu, P. Matsudaira, J. C. Rajapakse “Quantification of Cytoskeletal Protein Localization from High-Content Images”, WABOIOS 1st Anniversary Symposium 2010 (Poster presentation).

S. Zhu, P. Matsudaira, P. T. C. So, J. C. Rajapakse “Quantification of Cytoskeletal Protein Localization from High-Content Images”, BIRC Student Symposium 2010, Singapore (Oral presentation).

S. Zhu, P. Matsudaira, P. T. C. So, J. C. Rajapakse “Quantification of Cytoskeletal Protein Localization from High-Content Images”, SMA symposium 2011 (Poster presentation).

S. Zhu, C.K. Kwoh, R. Welsch, P. Matsudaira “Co-localization Quantification of Transcriptional Factors from 3D Images”, APOCB 2014 (Poster presentation)

Table of Contents

Acknowledgment	i
Publications	ii
Table of Contents	iii
Abstract	vi
List of Tables	vii
List of Figures	viii
Abbreviations	xiii
Chapter 1	1
Introduction	1
1.1 Motivation	1
1.1.1 Subcellular localization of protein is crucial for understanding protein functions and interactions	1
1.1.2 Studies of qualitative prediction of subcellular localization	3
1.1.3 Methods of quantification of co-localization	5
1.2 Objectives	7
1.2.1 Objective 1: Quantitative analysis of protein subcellular localization with co-localization features	7
1.2.2 Objective 2: Evaluation of the performance of co-localization features in prediction of subcellular localization with protein sequence features	8
1.2.3 Objective 3: Interpretable analysis of co-localization from high content 3D images	9
1.3 Contributions of this thesis	10
Chapter 2	12
Subcellular Localization Quantification with Co-Localization Features	12
2.1 Introduction	12
2.2 Computational Framework	15
2.2.1 Cell Segmentation	15
2.2.2 Intensity Transformation of Subcellular Compartments	19

2.2.3	Protein Segmentation	22
2.2.4	Co-Localization Quantification	26
2.2.5	Protein Localization Profiling.....	26
2.3	Experiments and Results	27
2.3.1	Sample Preparation	27
2.3.2	Imaging.....	27
2.3.3	Image Processing	29
2.3.4	Cell Selection.....	29
2.3.5	Co-localization Indexing	32
2.3.6	Protein Clustering	32
2.4	Conclusion.....	35
Chapter 3		36
Subcellular Localization Prediction with Co-Localization Features and Protein Sequence Features.....		36
3.1	Introduction	36
3.2	Feature Sets	37
3.2.1	Co-localization Features.....	37
3.2.2	Protein Sequence Features.....	40
3.3	Localization Classification	51
3.3.1	Ground Truth and Training Set.....	52
3.3.2	Bootstrapping.....	52
3.3.3	Class Binarization	53
3.3.4	Class Balancing.....	53
3.3.5	Feature Ranking	54
3.3.6	Selection of the Best Feature Subset.....	55
3.3.7	Classifier Training and Evaluation	56
3.4	Experiments and Discussion.....	57
3.4.1	Subcellular Localization Prediction with Co-Localization Features	57
3.4.2	Subcellular Localization Prediction with Co-Localization Features and sequence features	66
3.5	Conclusion.....	72

Chapter 4.....	73
Co-Localization from 3D Images.....	73
4.1 Introduction	73
4.2 Performance of Co-Localization Measurements	74
4.2.1 Synthetic Images	74
4.2.2 Co-localization measurements	78
4.2.3 Performance	81
4.3 Co-localization of Transcription Factors	85
4.3.1 Transcription Factors	85
4.3.2 Sample Preparation	95
4.3.3 Microscopic system and setting	99
4.3.4 Image Processing	101
4.3.5 Results and Discussion.....	102
4.4 Conclusion.....	120
Chapter 5.....	121
Conclusion and Future Directions.....	121
5.1 Conclusion.....	121
5.2 Future Work and Directions	122
Appendix.....	125
References.....	140

Abstract

The knowledge of subcellular localization and co-localization of proteins is crucial for investigating how proteins function and interact with each other within a cell. Plenty of research groups have dedicated their efforts in characterizing and predicting the subcellular localizations of proteins. In this thesis, I introduce the fluorescence microscopic term – co-localization into the process of quantification and prediction of subcellular localization, and the corresponding computational frameworks are developed to perform and evaluate the performance of the co-localization in quantifying and predicting of subcellular localization. The co-localization measurement can either work as a single parameter to estimate the relationship between a specific protein and a series of subcellular compartments to generate a co-localization profile; or work with many other statistic measurements serving as a set of co-localization features, which can be used to predict the subcellular localization itself or with other protein sequence features to improve the prediction accuracy. On the other hand, tested with synthetic images, we developed a method based on the co-localization measurement to estimate the co-occurrence and correlation separately, which would be greatly helpful in providing biological meaningful explanations to the quantification results, especially for mid-value results. The methods were validated and applied on 2-D cytoskeletal protein image dataset and 3-D transcription factor image dataset.

List of Tables

Table 1: Construct Map of Cytoskeletal proteins (Full name see Appendix I)	28
Table 2: Representative samples of cells transfected with 96 constructs in the construct map	28
Table 3. Co-localization of Protein Clusters	33
Table 4. Proteins in Protein Clusters	34
Table 5. Co-localization Features.....	38
Table 6 Amino Acid attributes and the Division of the Amino Acids	44
Table 7 Co-localization Features.....	49
Table 8 Sequence Features.....	50
Table 9. Comparison of Predicted Localization and Possible Functions	64
Table 10 Accuracy of four classes from feature merging method	68
Table 11 Accuracy of Actin and CytoS classes from Vote merging method	71

List of Figures

Figure 1 Computational Framework for subcellular localization quantification with co-localization measurement	14
Figure 2. A single representative image field example and the Comparison of three cell segmentation algorithms. (A) One single image field of HeLa cells with CORO2B protein transfected; the blue channel (B) is nuclei stained with Hoechst 33342, while red (C) is actin stained with Texas red - phalloidin, and green (D) is GFP-tagged transfected protein CORO2B. The transfection efficiency of CORO2B is low, as there are very few cells are transfected with relatively high intensity in this field example (D Insertion I and III), and some apoptotic cells with high intensity (D insertion II) should be removed before co-localization analysis. Regions (E) and (F) are zoomed in for segmentation comparison. The region growing method with multi-resolution segmentation (E1, F1) provided relatively more accurate segmentation results with smoother cell borders and less over- and under-segmentation, compared to Level set method with shape marker and marking function combined with nuclear information (E2,F2)and level set method with topological dependence (E3, F3).....	18
Figure 3. Illustration of subcellular compartments of a single representative cell example. (A)The raw image of a single cell transfected with WASPIP protein. Based on the method described in Section 2.2.3, the masks and the 3D illustrations of the intensity transformation for the five subcellular compartments are shown separately: Nucleus (C, D), Cytoplasm (E, F), Actin (G, H), Plasma Membrane (I, J), and Cytosol (K, L).The x and y axis label the dimensions of the cell image, and z axis labels the value of current pixel position in the intensity transformation (B).	21

Figure 4 Cell selection based on Nucleus size. Scatterplot shows the histogram of nucleus size of each segmented cells. Gaussian fit for the first peak is shown in red curve.	31
Figure 5 Cell samples for different GFP intensity levels. Cell samples are taken from cell set with Cofilin1 (CFL1) protein transfected, and the GFP intensities of the cell samples are from 5 to 60. The cells with GFP levels higher than 50 would be identified as toxic or dead cells.	31
Figure 6 Cell selection based on GFP intensity	31
Figure 7 The Framework of the classifier training process. The whole framework can be separated into two main processes: Images processing and localization classification. ...	37
Figure 8 The framework of localization classification.....	51
Figure 9. Performance of the localization classifiers: (A) the percentage of features selected for each classifier. (B) The accuracy of each classifier.	58
Figure 10. The feature frequency of four classifiers during the 50 times of feature selection with random sub-sampling training sets. (A)Nucleus Classifier (B) Actin Classifier (C) Plasma Membrane Classifier (D) Cytosol Classifier.....	59
Figure 11. The proportion changes of features grouped with different measure methods: (A) weighted overlapping area, (B) intensity correlation quotient, (C) overlap coefficient, and (D) Pearson's correlation coefficient.	60
Figure 12. The proportion changes of features grouped with different subcellular compartments: (A) Nucleus classifier, (B) Actin classifier, (C) Plasma Membrane classifier, and (D) Cytosol classifier.	62

Figure 13 Flow Path for the Feature merging approach. In this approach, the sequence features and image features are merged first before the following training steps.	67
Figure 14 Flow Path for the Vote merging approach. In this approach, the SVMs were trained separately with sequence features and image features. And the final prediction result is merged from two SVM flow.	70
Figure 15 The illustration of co-occurrence. In the overlay of red and green channel, there are both pixels occupied same and not same locations. Only the yellow rectangle in the middle of the image is 100% co-occurrence.	75
Figure 16 A representative sample of Synthetic Image with 30% co-occurrence and 0.6 Correlation. Img2' is the synthetic image generated from Img2 to have 30% co-occurrence and 0.6 correlation with Img1.	75
Figure 17 The illustration of Intensity Correlation. Given red channel, three green channel images are shown to demonstrate positive correlation, negative correlation and random correlation. Note for intensity correlation only the intensities of the central rectangle count, as other pixels are not even co-occupied the same location with red channel.	76
Figure 18 Illustration of Synthetic Images with variations of Correlation. The synthetic images are generated from Img2 to have correlation variations from 0.1 to 1.0 with Img1.	77
Figure 19 Performance of co-localization measurements on synthetic images generated by manipulating intensities.	81
Figure 20 Performance of co-localization measurements on synthetic images generated by mixing positive correlated pixels with random correlated in overlapping area.	83

Figure 21 Performance of co-localization measurements on synthetic images generated by mixing positive correlated pixels with negative correlated in overlapping area.....	84
Figure 22 heatmap of the binding profiles of 13 factors. factors were clustered on both axes based on their binding profiles. Colors in the heat map reflect the co-localization frequency of each pair of TFs in MTL (yellow means more frequently co-localized, red means less) (Chen, X et al. 2008).	94
Figure 23 flow profile of cell synchronization under different conditions. (A) Flow profile for control cell population. Two peaks were observed to show cell population in S and G2/M phases. (B) After treated with Thymidine, the cells are synchronized in S phase and the peak at G2/M phase is disappeared. (C) Treated with Nocodazole only. Cells were synchronized in G2/M phase. Peak for S phase disappeared. However, the cell survival rate decreased and the peak in G2/M phase was broad and will less cell number. (D) Double blocking with Thymidine followed by nocodazole. The cells are well synchronized in G2/M phase with a very sharp peak in the flow profile.	96
Figure 24 flow profile of cell releasing time points after synchronization. After double blocking, the drugs are washed out with PBS and cells are released from the blocking and went into the next cell cycle phase. After 3 hrs majority of cells went through the cell cycle and back to S phase.	97
Figure 25 Four time points chosen in the following analysis	98
Figure 26 3D projections of image samples of 9 TFs. A single representative colony shown for each factor.	99
Figure 27 Double staining for each transcription factor pairs – Nanog-centric cluster .	100

Figure 28 Double staining for each transcription factor pairs – between Nanog-centric cluster and Myc-centric cluster	100
Figure 29 Double staining for each transcription factor pairs –Myc-centric cluster	101
Figure 30 Co-localization heatmap of TF pairs in S phase.....	103
Figure 31 Co-localization heatmap of TF pairs in G2/M phase	103
Figure 32 Co-localization heatmap of TF pairs in M phase	104
Figure 33 Co-localization heatmap of TF pairs in G1 phase	104
Figure 34 Co-localization of Oct4.....	107
Figure 35 Co-localization of Sox2	108
Figure 36 Co-localization of Nanog	109
Figure 37 Co-localization of P300	112
Figure 38 Co-localization of CBP	113
Figure 39 Co-localization of N-myc.....	116
Figure 40 Co-localization of C-myc.....	117
Figure 41 Co-localization of ZFX	118
Figure 42 Co-localization of E2F1	119

Abbreviations

2D	Two dimension
3D	Three dimension
APAAC	Amphiphilic pseudo amino acid composition
AUROC	The area under the Receiver-operating-characteristics curve
CBP	CREB-binding protein
CTD	Composition, transition, distribution
CytoP	Cytoplasm
CytoS	Cytosol
DBD	DNA-binding domain
DMEM	Dulbecco's Modified Eagle Medium
DNA	Deoxyribonucleic acid
ES cell	Embryonic stem cell
FBS	Fetal bovine serum
FCS	Fluorescence correlation spectroscopy
FRAP	Fluorescence recovery after photo-bleaching
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase
GFP	Green fluorescent protein
HeLa	Cell line derived from human cervical carcinoma cells, from Henrietta Lacks the female patient
hESC	Human embryonic stem cell
HTH	Helix-turn-helix
ICQ	Intensity correlation quotient

iPSC	Induced pluripotent stem cell
LIF	Leukemia inhibitory factor
mESC	Mouse embryonic stem cell
MOC	Manders' Overlap Coefficient
MOCp	Manders' overlap coefficient for protein channel
MRF	Markov random fields
Nuc	Nucleus
OCT4	Octamer-binding transcription factor 4
OOB	Out-of-bag
P300	E1A binding protein p300
PBS	Phosphate Buffered Saline
PCC	Pearson's Correlation Coefficient
PM	Plasma membrane
POU	PIT/OCT/UNC
QSO	Quasi-sequence-order
RLSM	Reflected light-sheet microscopy
RNA	Ribonuclei acid
SOX2	SRY (sex determining region Y)-box 2
SRC	Spearman's rank correlation coefficient
SSD	Signal-sensing domains
SVM	Support Vector Machine
SVM-RFE	Support vector machine recursive feature elimination
TAAP	Total amino acid property
TAD	Transactivation domain

TF	Transcription factor
TIRF	Total internal reflection fluorescence microscopy
UNIPROT	Universal Protein Resource knowledgebase
WOA	Weighted overlapping area
ZFX	Zinc finger protein, X-linked

Chapter 1

Introduction

1.1 Motivation

1.1.1 Subcellular localization of protein is crucial for understanding protein functions and interactions

Although there are almost 100 million different living species on Earth, cells are the most basic and smallest structural and functional units of all live organisms. Prokaryotes, usually single-celled organisms, consist of a single intracellular compartment surrounded by a plasma membrane; while eukaryotic cells have much more complex structures: A set of functionally distinct, lipid membrane-enclosed, well organized compartments are surrounded and protected by a plasma membrane (Alberts, Bray et al. 1994). Those subcellular compartments, also called organelles, have their own set of enzymes and specialized molecules to ensure their characteristic structural and functional properties (Blobel 1980). With those subcellular compartments, specific biomolecules can be contained and restricted in a certain region, thus to establish physical boundaries for biological processes and therefore different metabolic activities can be carried out at the same time in one cell.

Common subcellular compartments include: Nucleus, Endoplasmic reticulum, Mitochondria, Golgi apparatus, Plasma membrane, Cytosol, Microtubule, etc. Each subcellular compartment has its specific functions and properties, which are distinct from others. For example, the nucleus containing the main genome is where DNA and RNA

synthesis is performed. The cytosol is the location for protein synthesis and degradation. The plasma membrane separates the interior of the cell and thus protect it from the outside environment, and is crucial for cellular processes including cell signaling, cell adhesion and cell movement. The rough endoplasmic reticulum is the place of protein synthesis with ribosomes bounded on the cytosolic face. The cytoskeleton including microfilaments, intermediate filaments and microtubules help maintain cell shape, organize cytoplasm, supporting cellular motility and intracellular transporting (Frixione 2000, Damania, Subramanian et al. 2010). In order to actualize all those distinct functions of different subcellular compartments, the coordinate efforts of specific sets of proteins are required.

Most proteins are synthesized on ribosomes in the cytosol, few on the ribosomes of mitochondria and plastids. Each protein has its unique amino acid sequence, from which the protein is encoded and assembled. The amino acid sequence of a protein also contributes to its destiny and sorting signals will direct delivery to subcellular locations. There they will perform their functions and actualize the characteristic structural and functional properties of each subcellular compartment (Blobel 1980, Palade 1992). Only when the mature proteins are transported to the corresponding subcellular structures, the biological functions can be exerted. Deviation from protein localization will lead to serious function loss or dysfunction in the organism. Abnormal protein subcellular localization is discovered in various diseased cells, such as cancer cells. Protein subcellular location is one of the most important properties of protein, which helps to predict proteins functions, reveal the mechanism of molecular interaction, and understand

the complex physiological process making protein subcellular localization research of great significance in cell biology, proteomics, and drug design research.

1.1.2 Studies of qualitative prediction of subcellular localization

Many research groups have characterized and predicted the subcellular localizations of proteins through both biochemical experiment methods and bioinformatics methods. Traditional biochemical experiment methods such as cell fractionation, electron microscopy and fluorescence microscopy detect subcellular localization accurately by direct visualization, however, those methods are time-consuming and costly in dealing with large amounts of target proteins. The development of bioinformatics with protein sequences provided a stage for machine learning methods in this field. First in 1994, were attempts to discriminate intra- and extracellular proteins with amino acid composition and residue-pair frequencies (Hiroshi, Nakashima et al. 1994). The first protein subcellular localization prediction system based on support vector machine was built in 2001 (Hua and Sun 2001). In the past 20 years, a variety of prediction methods with machine learning methods have emerged based on different feature extraction, such as amino acid sequences (Hua and Sun 2001, Qiu, Xiao et al. 2013), sorting signals (Claros, Brunak et al. 1997, Nielsen, Engelbrecht et al. 1997, Zhang et al. 2008), physicochemical properties (Sarda, Chua et al. 2005), primary protein structures (Han, Anh et al. 2013), substantiality motif (Tang, Li et al. 2013), position-specific scoring matrix (Lin et al. 2013), protein-protein interaction (K, Lee, et al. 2008), etc.

On the other hand, different from traditional biochemical methods, image features extracted by using fluorescent microscopy and image processing techniques have been

frequently used with machine learning methods to identify subcellular locations of proteins (Boland, Markey et al. 1998, Hu and Murphy 2004, Huang and Murphy 2004, Farkash-Amar, Eden et al. 2012). Compared to the bioinformatics methods, the image features provide direct information of subcellular locations from the protein images, thus yield the importance of the quality of the image training set. Well-established training sets are established for a standard comparison between different prediction methods. Different machine learning methods are employed in the field for prediction of protein subcellular localization, such as Neural Network Classifier (Boland and Murphy 2001), Markov random fields (MRF) (Murphy, Weiss et al. 1999), Bayesian networks (Boland and Murphy 2001), graph theory model (Xu, Lu et al. 2013), Support Vector Machine (Hua and Sun 2001, Chen, Zhao et al. 2007), etc. The great power of these methods is shown through these studies with prediction accuracy reaching 85%. However, most of these methods are intended for proteins with single and fixed subcellular localization pattern.

How different proteins interact with more than one subcellular compartment and their presence in more than one location has not been addressed. Many cellular compartments need to be considered as predicting candidates for a protein's location and prediction of only single subcellular protein pattern is not sufficient. Several methods have been presented to address this problem as supervised pattern un-mixing (Zhao, Velliste et al. 2005, Peng, Bonamy et al. 2010) and unsupervised pattern un-mixing (Coelho, Peng et al. 2010). For the supervised pattern un-mixing methods, identification of fundamental patterns will become a difficult task due to the variety of experimental conditions and imaging techniques. The reported unsupervised pattern un-mixing method defined a set

of features to describe the localization patterns of proteins. The method claimed to to un-mix two different patterns and quantify the relative amount of proteins in two organelles. However, it still is difficult to localize proteins in more than two subcellular compartments.

Thus there is a need to establish a method to identify and quantify the dynamic localizations of proteins with multiple subcellular compartments. A more quantitative definition of the term, co-localization, is introduced into this field and is the focus of this thesis.

1.1.3 Methods of quantification of co-localization

Before quantifying and analyzing co-localization, the limitations of the microscopic system need to be aware. The objective lens only allows partial collection of light and the parameter – numerical aperture (NA) – is used to quantified this capability. NA determine the ability to distinguish between two adjacent punctate light sources. The minimal distance between the two Airy discs of the light sources defines the optical resolution or the separating power. Biological samples are usually not two-dimensional limited, thus the diffraction pattern of light should be considered as 3D information and thus to define the point spread function (PSF). Therefore, it is important to note that when doing the image acquisition for co-localization analysis, several subsequent optical sections must be carefully considered such as in two or three dimensions, near to the resolution limit of the optical system, chose the appropriate justified pixel size in both xy and z direction. Down-sampling and over sampling need to be aware during the process.

Co-localization is the fluorescence microscopic term to measure the spatial overlap

between two different fluorescent labels. The co-localization measurements can be grouped into two categories: pixel-based/correlation coefficient-based approaches and object-based approaches. In pixel-based co-localization quantification, a series of global statistical measurements are mainly used to assess the relationship between fluorescence intensities in different compartments. Pearson's coefficient (Manders, Stap et al. 1992) is the most popular and simple measurement to quantify the linear approximation from the linear regression of two fluorochromes. Manders' overlap coefficient (Manders, Stap et al. 1992) is a measurement based on the Pearson's correlation coefficient without the average intensity terms in the mathematical expression. Spearman correlation coefficient (Adler, Pagakis et al. 2008) is another measurement based on the Pearson's correlation coefficient using the rank value instead of absolute intensity value in the mathematical expression. Costes' approach (Costes, Daelemans et al. 2004) is another statistical significance algorithm based on Pearson's coefficient which including a pre-step of automatic thresh holding to eliminate noise. Additional quantification methods of co-localization are intensity correlation analysis (Li, Lau et al. 2004) and rank-based weighting coefficient (Singan, Jones et al. 2011). These correlation measurements are mathematically focused on different aspects of the co-localization: some are sensitive indicators for the linear relationships between intensity distributions of channels, while others are good at reporting the co-occupied pixels.

In the studies of object-based co-localization quantification, instead of only relying on individual pixel coincidence, the unique structures of the object of interest are also taken into account. The two most common methods are comparing three-dimensional centroids (Lachmanovich, Shvartsman et al. 2003, Boutté, Crosnier et al. 2006) and object image

combining intensity correlation coefficient-based analysis (Jaskolski, Mulle et al. 2005). The advantage of object-based co-localization compared to pixel-based methods is by pre-defining the region of interest before actually quantify the correlation between two image channels, it will greatly reduce the bias generated by the background noise and unspecific signal; and by taking the structures of the object of interest into account, the analysis will gain a much more accurate and specific results to the certain research topics.

1.2 Objectives

1.2.1 Objective 1: Quantitative analysis of protein subcellular localization with co-localization features

A large number of papers identify and predict the localization patterns of proteins from fluorescence microscopy images with machine learning methods, however, most of those methods are intended for proteins with a single and fixed subcellular localization pattern. How different proteins interact with more than one subcellular compartment and their presence in more than one location has not been addressed. More than one cellular compartment needs to be considered in predicting candidates for a protein's location and prediction of only single subcellular protein pattern is not sufficient.

Our aim is to introduce co-localization – a fluorescence microscopic term to measure the spatial overlap between two different fluorescent labels – in the process of quantifying protein subcellular localization. Instead of measuring the co-localization between two proteins, the co-localization concept is introduced and quantified to measure the co-localization between proteins and subcellular compartments. Quantifying a single co-

localization parameter is necessary so that a variety of proteins can be quickly and easily compared without bias. By quantify the co-localization of proteins with a set of subcellular compartments, the co-localization profile for each protein can be generated and thus provide a clear co-localization pattern of highly dynamic proteins. Proteins with unknown functions can be investigated by comparing with co-localization profiles generated in a cytoskeletal protein library.

1.2.2 Objective 2: Evaluation of the performance of co-localization features in prediction of subcellular localization with protein sequence features

In the studies of pixel based co-localization quantification, a series of global statistical measurements are mainly used to assess the relationship between fluorescence intensities in different compartments, including Pearson's coefficient overlap coefficient, a statistical significance algorithm based on Pearson's coefficient, intensity correlation analysis, rank-based weighting coefficient, etc. Those correlation measurements describe different aspects of the co-localization: some are sensitive indicators for the linear correlation between channels, while some are good at reporting co-occupying of pixels. Hence, a certain set of correlation measurements would be able to serve as co-localization features for prediction of subcellular localization. On the other hand, well-developed computational methods are reported for subcellular localization prediction purpose based on the sequence information, such as amino acid composition, dipeptide composition, and gapped amino acid pairs, as well as hydrophilicity of amino acids, functional domain composition and psi-blast profile.

We aim is to predict the subcellular localization of cytoskeletal proteins with a set of co-localization features, to investigate the performance of these co-localization features in the trained classifiers. Furthermore, we aim to combine co-localization features and protein sequence features in the process of subcellular localization prediction to prove the co-localization features would be good complementarities to the protein sequence features to improve the prediction accuracy.

1.2.3 Objective 3: Interpretable analysis of co-localization from high content 3D images

Quantification of co-localization is the combination of both co-occurrence and intensity correlation. Co-occurrence describes the co-occupying events of pixels; while intensity correlation describes the linear relationship between intensity distributions. Since almost all the co-localization measurements are sensitive to both signal co-occurrence and intensity correlation, it become very difficult to interpret mid-value of co-localization.

We aim to propose a method to separate the aspects of co-occurrence and correlation from the co-localization quantification, which would be greatly helpful in providing biological meaningful explanations to the quantification results. And we aim to apply this method in quantifying the co-localization of transcriptional factor pairs to analyze the co-localization between these factors in order to investigate the correlation and interaction between factors and also the comparison of the co-localization of factor pairs within control cells and treated cells.

1.3 Contributions of this thesis

In this thesis work, the fluorescence microscopic term co-localization is quantified, evaluated and interpreted in the processes of quantifying subcellular localization of a protein, predicting subcellular localization from quantification measurements and quantifying co-localization between two objects. The research in this thesis has made three main specific contributions:

1. A computational framework to quantify the co-localization between a protein and a subcellular compartment.

The fluorescence microscopic term – co-localization was introduced into the process of quantification of subcellular localization. We proposed a new co-localization measurement to meet the requirement of this certain problem and the corresponding high throughput computational frameworks are developed to perform and evaluate the performance of the co-localization in quantifying of subcellular localization. The co-localization profiles can be generated for highly dynamic proteins to address and monitor their dynamic localization within the cell. Proteins with unknown functions can be compared the localization patterns with existing protein clusters to gain an insight of its possible similarities in localizations and functions with other proteins.

2. A good complementary feature set for prediction of protein subcellular localization.

The co-localization measurements can serve as a set of co-localization features, which can be used to predict the subcellular localization itself or work with protein sequence features to improve the prediction accuracy.

3. An interpretable method of co-localization quantification by separating the aspects of co-occurrence and intensity correlation from the co-localization event.

After testing synthetic images, we developed a method based on the co-localization measurement to estimate the co-occurrence and correlation separately, which would be greatly helpful in providing biological meaningful explanations to the quantification results, especially for mid-value results. And the method first allows the visualization of co-occurrence pixels on the pseudo-color image to provide a direct idea where and how much two proteins co-localized in the area of interesting.

Chapter 2

Subcellular Localization Quantification with Co-Localization Features

Part of this chapter is published as:

Zhu, S., P. Matsudaira, R. Welsch and J. C. Rajapakse (2010). Quantification of cytoskeletal protein localization from high-content images. *Pattern Recognition in Bioinformatics*, Springer: 289-300.

2.1 Introduction

The cytoskeleton, a complex network structure in all eukaryotic cells and prokaryotic cells, is a cellular skeleton that is involved in many important cellular processes, such as establishing cell shape, organizing cytoplasm, supporting cellular machinery for motility, intracellular transporting of organelles, organizing cells into tissues and signaling (Frixione 2000, Damania, Subramanian et al. 2010). There are three main types of cytoskeletal filaments contained in eukaryotic cells, which are microfilaments, intermediate filaments, and microtubules (Alberts 2008). Those three types of filaments have distinct mechanical properties and dynamics, assembled with three families of protein families. Besides, more cytoskeletal proteins associate with and are involved in regulating the formation, turnover of the filaments, and also serve as tracking protein, nucleus, linker, motor, binding protein, capper, depolymerizing protein, and many other roles functioning together with three main filaments to implement all those cellular processes (Pollard and Borisy 2003, Moseley 2013). Signaling through a complex

functional network of small molecules and enzymes ensure the communication and connection of the cytoskeletal proteins to the metabolic reactions. With those signals, cytoskeletal proteins respond and function harmoniously to the coordinated efforts of cellular networks. (Bray 2001, Khurana 2006). However, the mechanism of the sophisticated coordinated system of these proteins is still not fully revealed. Therefore, investigation of the exact functions and interactions of the whole cytoskeletal protein library would reveal the mechanisms of cellular mechanics, cellular dynamics, metabolic signaling pathway modeling, subcellular organization, early development of cancer, etc (Janmey 1998, Leonardy, Freymark et al. 2007, Hung and Link 2011, Farkash-Amar, Eden et al. 2012). Due to the dynamic characteristic of the cytoskeletal proteins, they were chosen as experiment data in this chapter to investigate the performance of my quantification method of protein subcellular localization.

Fluorescent microscopy identifies protein subcellular locations through image processing, feature extraction, and pattern recognition(Hu and Murphy 2004, Huang and Murphy 2004). Machine learning methods have been previously applied for identifying and predicting the localization patterns of proteins by using training data. But such methods were intended for single cell and usually a single predicting subcellular compartment.

How different proteins interact with more than one subcellular compartment and their presence in more than one location has not been addressed. More cellular compartments need to be considered as predicting candidates for a protein's location and prediction of a single subcellular protein pattern is not sufficient. Thus I chose to focus on fluorescent signal co-localization – a measurement of overlap between two signals. Quantifying a single co-localization parameter is necessary so that a variety of proteins can be quickly

and easily compared without bias. Determining the co-localization between cytoskeletal proteins and subcellular components—such as nucleus, cytoplasm, plasma membrane, actin network, etc., will help to define and simplify the proteins’ operations and locations. In the protein subcellular co-localization analysis, the intensities within subcellular compartments are not a major concern. For example, DNA is stained to represent the nucleus. The DNA intensities in nuclear compartment may not be uniform, but our major concern is how the protein is co-localized in the nucleus.

Therefore, in this chapter, I will introduce a new co-localization measurement generating co-localization profiles indicating the amount of co-localization of proteins and subcellular compartments. And a computational framework is developed and described in this chapter to actualize the whole image processing procedure.

Figure 1 Computational Framework for subcellular localization quantification with co-localization measurement

2.2 Computational Framework

In order to quantify subcellular localization of cytoskeletal proteins with a single parameter, the co-localization, I developed a computational framework involving individual cell segmentation, protein segmentation, intensity transformation of subcellular compartments, and co-localization quantification (Fig. 1). In what follows, I will describe the different steps involved in our approach.

2.2.1 Cell Segmentation

Images were segmented into small objects using a multi-resolution segmentation technique based on object-oriented image analysis. This method is used to create object primitives as the first processing step in the segmentation analysis. The criterion for the segmentation is that average heterogeneity of image objects weighted by their size in pixels should be minimized. The merging procedure starts with each voxel forming one image object or region. For possible merging the degree of fitting is evaluated and the merge is fulfilled if it is smaller than a given least degree of fitting. The procedure stops when there are no more possible merges. The definition of degree of fitting is:

$h_{diff} = \sum_c w_c (n_1(h_{mc} - h_{1c}) + n_2(h_{mc} - h_{2c}))$, where c denotes the channel, and w_c is the

weight of the channel. n_1 and n_2 represent the size of object 1 and 2, respectively. h_{1c} , h_{2c} , and h_{mc} represent the heterogeneity of object 1, 2, and merged object of object 1 and 2, respectively.

The definition for spectral heterogeneity of image objects is the variance of spectral mean values or the standard deviation of spectral mean values. After the merging, the whole image is segmented into a set of image objects.

I assign each image object a numerical label. Now the images are projected to a subset in one-dimensional label space $\Gamma = \{1, 2, \dots, N\}$. N is the total number of image objects. Let

$f_n : \Omega \rightarrow R$ and $f_a : \Omega \rightarrow R$ represent the images of nuclear channel and actin channel,

respectively, where $\Omega \subset R^2$ is the 2D image domain. Let $g_n : \Gamma \rightarrow R$ and $g_a : \Gamma \rightarrow R$ represent the average object intensities of nuclear channel and actin channel, respectively.

Let O_l represent the set of voxels belonging to the object l , $g_n(l)$ and $g_a(l)$ represent the average intensities of nuclear channel and actin channel of object l , respectively. $N(l)$ is the set of objects in the neighborhood of object l .

After the primary segmentation, the image objects are classified as nuclear objects and cell body objects based on flexible thresholds of nuclei and cell intensities. The nucleus objects are used as seeds for region growing method for cell segmentation, on the assumption that each cell has only one nucleus (Zhu, Matsudaira et al. 2010). A rule set was then developed for cell segmentation. The cell body object was fused with its

neighbor nucleus. When one cell body object has more than one neighbor nuclear object, it was fused with the nuclear object that shared the largest border with it. The region growing method with multi-resolution segmentation This new method provided better segmentation results than other advanced segmentation algorithms (Fig. 2E, F): Level set method with shape marker and marking function (Cheng and Rajapakse 2009) combined with nuclear information and level set method with topological dependence (Yu, Lee et al. 2009). Level set segmentation is an important image segmentation techniques based on partial differential equations. Progressive evaluation on the differences will be performed among the neighboring pixels to find the object boundaries, and the algorithm will converge at the boundary of the object where meet the highest differences. Cheng et al. has introduced shape markers extracted with an adaptive H-minima transform and outer distance transform based marking function in the level set algorithm to accurately separate clustered nuclei. And Yu et al. introduced dynamic watershed lines based on the constraint of topological dependence in level set method to merge split cells.

In the experiments, cells were labeled with fluorescence to highlight nuclei, actin, and proteins of the cells. The green fluoresce protein (GFP) was selected to identify a particular protein. The nuclear and actin channels were used to identify subcellular compartments. Let $f_n : \Omega \rightarrow R$, $f_a : \Omega \rightarrow R$ and $f_g : \Omega \rightarrow R$ represent the images of nuclear channel, the actin channel, and the GFP channel, respectively, where $\Omega \subset R^2$ is the 2D image domain and $x \in \Omega$ denotes the 2-D coordinates of a pixel site in the image. Let $w_i \subset \Omega$ denote the area of an individual cell, and $i \in N$ be the labels of cells in one image.

Figure 2. A single representative image field example and the Comparison of three cell segmentation algorithms. (A) One single image field of HeLa cells with CORO2B protein transfected; the blue channel (B) is nuclei stained with Hoechst 33342, while red (C) is actin stained with Texas red - phalloidin, and green (D) is GFP-tagged transfected protein CORO2B. The transfection efficiency of CORO2B is low, as there are very few cells are transfected with relatively high intensity in this field example (D Insertion I and III), and some apoptotic cells with high intensity (D insertion II) should be removed before co-localization analysis. Regions (E) and (F) are zoomed in for segmentation comparison. The region growing method with multi-resolution segmentation (E1, F1) provided relatively more accurate segmentation results with smoother cell borders and less over- and under-segmentation, compared to Level set method with shape marker and marking function combined with nuclear information (E2, F2) and level set method with topological dependence (E3, F3).

2.2.2 Intensity Transformation of Subcellular Compartments

In this study, five subcellular compartments were studied: nucleus, cytoplasm, actin (as we focused on cytoskeletal proteins), plasma membrane, and cytosol. For most of the subcellular compartments, the intensity distributions of labeled signals were not uniform, but the co-localization amounts of proteins were different. Thus the intensity transformations of subcellular compartments are performed using intensity information of actin and nucleus staining as well as the position and relation information of subcellular compartments. Instead of computing the co-localization of the protein signal and the compartment signal, I compute the co-localization of the protein signal and the intensity transformed images of the compartments. (Zhu, Matsudaira et al. 2010).

Let a compartment be denoted by $c = \{nucleus, cytoplasm, actin, membrane, cytosol\}$, and

$\tilde{f}_c : \Omega \rightarrow R$ denote the intensity transformed image in the compartment c (Fig. 3).

Nucleus: This is the compartment identified from the nuclei channel. I keep the intensities of the nucleus compartment uniform (Fig. 3 C, D).

$$\tilde{f}_{nucleus}(x) = \begin{cases} 1, & f_n(x) \geq t_n \\ 0, & f_n(x) < t_n \end{cases}, \quad x \in w_i; \quad (1)$$

t_n represents the threshold of the nuclei channel. The threshold was generated automatically with Otsu's method.

Cytoplasm: This is the compartment identified as the non-nuclear region. The transformation for the cytoplasm compartment is kept uniform (Fig. 3 E, F).

$$\tilde{f}_{cytoplasm}(x) = 1 - \tilde{f}_{nucleus}(x), \quad x \in w_i; \quad (2)$$

Actin: This is the compartment identified from the actin channel. The actin network warps the nucleus, so there would be an overlap between the nucleus and actin compartments. A ceiling threshold is used to account for over-saturation and the intensity values in the actin channel were rescaled to the range [0, 1] (Fig. 3 G, H).

$$\tilde{f}_{actin}(x) = \begin{cases} 1, & f_a(x) \geq t_a; \\ \frac{(f_a(x) - t'_a)}{(t_a - t'_a)}, & t_a > f_a(x) \geq t'_a; \\ 0, & f_a(x) < t'_a. \end{cases} \quad x \in w_i; \quad (3)$$

where t_a and t'_a are positive numbers representing upper and lower thresholds of the actin

channel. In this analysis, $t_a = 90\% \cdot \max(f_a(x))$ and $t'_a = 10\% \cdot \max(f_a(x))$

Plasma Membrane: This is the compartment identified as the border region of a cell. The intensities of plasma membrane were transformed using an exponential function of the minimum Euclidean distance to the cell border (Fig. 3 I, J).

$$\tilde{f}_{membrane}(x) = t_m \cdot \exp \left[- \left(\frac{(d(x) - t'_m)}{t''_m} \right)^2 \right], \quad x \in w_i; \quad (4)$$

Where $d(x) = \min \sqrt{(x - x')^2}, \quad x \in w_i, x' \in \partial w_i; \quad (5)$

and t_m, t'_m, t''_m are positive numbers representing the parameters of the exponential function with ∂w_i representing the border. The fitting parameters for the exponential function in the experiment are: $t_m=1.054$, $t'_m=1.53$, $t''_m=2.241$.

Cytosol: This is the compartment identified as cytoplasm without the components of actin and the plasma membrane (Fig. 3 K, L).

$$\tilde{f}_{cytosol}(x) = \max\{0, \tilde{f}_{cytoplasm}(x) - \tilde{f}_{membrane}(x) - \tilde{f}_{actin}(x)\}, \quad x \in w_i; \quad (6)$$

Figure 3. Illustration of subcellular compartments of a single representative cell example. (A) The raw image of a single cell transfected with WASPIP protein. Based on the method described in Section 2.2.3, the masks and the 3D illustrations of the intensity transformation for the five subcellular compartments are shown separately: Nucleus (C, D), Cytoplasm (E, F), Actin (G, H), Plasma Membrane (I, J), and Cytosol (K, L). The x and y axis label the dimensions of the cell image, and z axis labels the value of current pixel position in the intensity transformation (B).

2.2.3 Protein Segmentation

In order to identify the protein localization, cells were transfected with the protein tagged with GFP. By localizing the scattering of GFP-tagged proteins in the cells, its localizations in different subcellular compartments were identified. Before identifying the subcellular localization of the protein, the segmentation of protein needs to be correctly performed. Since the GFP intensities vary in different cells and compartments in different images, protein segmentation becomes a crucial component in finding the balance between capturing most of the protein information and highlighting the most specific protein information. Because the protein exists as small units, it cannot be segmented into one connected component. Thus, classical segmentation algorithms, such as watershed and region growing, become unsuitable. An automated thresholding method of identification of co-localized pixels has been earlier developed for protein-protein co-localization analysis (Costes, Daelemans et al. 2004).

Therefore, I developed an algorithm to segment the proteins by thresholding based on the correlation of its intensities with that of the responding compartment. The basic idea is to preserve most of the GFP pixels that are correlated with the intensities of the cell or its compartments, and remove the pixels that are least correlated or distributed almost randomly.

For each candidate threshold, the correlation coefficient is computed on both selected (intensities higher than the candidate threshold) and unselected pixels (intensities lower than the candidate threshold). Correlation coefficients high on selected pixels and low on unselected pixels indicate that the current thresholds can save high correlated pixels and

remove low correlated pixels, respectively. Thus, I want to achieve the proper balance between these two thresholds.

Several correlation coefficients were tested such as Pearson correlation coefficient and overlap coefficient. The intensity correlation quotient (ICQ) provided the best segmentation results. For one given cell image with channel f_1, f_2 , and a given image region W , the ICQ value is based on the intensity correlation coefficient (Li, Lau et al. 2004). The correlation coefficient at a pixel is:

$$\rho(f_1, f_2, x) = (f_1(x) - \mu_1)(f_2(x) - \mu_2), \quad x \in W; \quad (7)$$

μ_1 and μ_2 denote the mean intensities values with the region W of the two channels.

The ICQ is defined as the ratio of the positively correlated pixels and the negatively correlated pixels in the region W . The correlation coefficient for two denoted areas is:

$$\rho(f_1, f_2, W) = \frac{\delta^+}{\delta^+ + \delta^-} - 0.5, \quad (8)$$

where $\delta^+ = \sum_{x \in W} \delta(\rho(x) > 0)$ donates the total number of positively correlated pixels, and

$\delta^- = \sum_{x \in W} \delta(\rho(x) < 0)$ donates the total number of negatively correlated pixels. The range

of ICQ falls between $[-0.5, 0.5]$. When $ICQ \approx 0$, random correlation; $-0.5 \leq ICQ < 0$, negative correlation; $0 < ICQ \leq 0.5$, positive correlation.

The Algorithm 1 gives a way to determine the optimum threshold for protein

segmentation within the cell. $f_{cell}:\Omega\rightarrow R$ is the cell image obtained by combining the nuclei channel and the actin channel: $f_{cell}(x)=f_a(x)+f_n(x), x\in w_i$. T is the final threshold generated for protein segmentation in this particular cell region w_i .

Algorithm 1 Determination of Threshold for Protein Segmentation

Input: $f_g(x)$ Intensity distribution of x on GFP channel

$f_{cell}(x)$ intensity distribution of x in cell region

Output: T the threshold

begin

$t = \max_x \{f_g(x) : x \in w_i\}$

$t_1 = t_2 = 0$

$r_1 = 0$

$r_2 = 1$

while $t \geq 0$

$W_{GFP} = \{x : f_g(x) \geq t, x \in w_i\}$

$\bar{W}_{GFP} = \{x : f_g(x) < t, x \in w_i\}$

if $\rho(f_{cell}, f_g, \bar{W}_{GFP}) < r_1$

$t_1 = t$

endif

if $\rho(f_{cell}, f_g, W_{GFP}) > r_2$

$t_2 = t$

endif

$t = t - 1$

endwhile

$T = \frac{t_1 + t_2}{2}$

end

2.2.4 Co-Localization Quantification

After segmentation of proteins within the cell and intensity transformation of subcellular compartments, the co-localization of proteins and subcellular compartments is quantified by a “co-localization” measurement that gives a better understanding about the percentage protein localized in a compartment. The co-localization of a protein p in a compartment c is defined as:

$$Coloc(c, p) = \frac{\sum_{x \in W_p} f_p(x) \cdot \tilde{f}_c(x)}{\sum_{x \in W_p} f_p(x)} ; \quad (9)$$

where W_p represents the set of pixels in the region occupied by the protein p . $f_p(x)$ is the intensity distribution of the GFP channel highlighting protein p and $\tilde{f}_c(x)$ is the intensity transformation of the compartment c .

2.2.5 Protein Localization Profiling

Using a library P of GFP-tagged cytoskeletal protein constructs, I compute the co-localization values of the protein in the five subcellular compartments. For given protein $p \in P$, the co-localization profile is $Col(p) = \{Coloc(c, p)\}$, where $c = \{nucleus, cytoplasm, actin, membrane, cytosol\}$, then those protein profiles are clustered in order to find the functional proteins.

2.3 Experiments and Results

2.3.1 Sample Preparation

The cell preparation and image set came from Stephanie Reed, Matsudaira lab, MIT.

HeLa (human cervical carcinoma) cell line was used as the cell model in the experiments.

The cells were maintained with Dubelcco's Modified Eagle Medium (DMEM) supplemented with 10% fetal bovine serum (FBS) and split into two 96-well plates. Cell density is 10,000 per well before transfection. Lipofectamine2000 transfections were taken to the 96 well plates with 89 Invitrogen GFP-tagged cytoskeletal protein constructs at concentration 10ng/ul (Table 1, 2, Appendix I). The cells were fixed and stained with Hoechst33342 (nuclei) and Texas red phalloidin (actin).

2.3.2 Imaging

Imaging of transfected cells was performed by Cellomics vHCS: Scan V Target Activation application system with 20X magnification. For each image sample there were 96 wells containing 48 constructs transfection results; and for each well there were 40 non-overlapping imaging fields being scanned. Thus, the number of images in this dataset is about 7680. For each image, there are three fluorescent channels: blue (Hoechst33342) staining nuclei, red (Texas red - phalloidin) staining actin, and green (GFP) staining the particular protein. Those channels are denoted as f_n , f_a and f_g for nuclear channel, the actin channel, and the GFP channel, respectively.

Table 1: Construct Map of Cytoskeletal proteins (Full name see Appendix I)

	1	2	3	4	5	6	7	8	9	10	11	12
A	DCAMKL1	CORO2B	PDLIM3	WASPIP	KARCA1	CTTN	VIL1	ACTB	WASL	DNM2	ZYX	TUBA6
B	CDC42	CFL1	TGOLN2	TAGLN	WAS	EVL	TUBE1	TNS	CLIPR-59	DSTN	ITGB2	CORO1B
C	ATP6V1C1	TAGLN3	VIL2	TUBD1	ATP5G2	TUBGCP3	PARVG	RDX	ITGB3BP	FSCN3	TEKT3	PAK4
D	GJB2	KALPHA1	CAPZA3	ACTG1	KPTN	TEKT1	PTK9	KIF3A	TPM1	TUBB2	ITGB1	ADAM15
E						PLS1	VASP	HTATIP	ARHGEF6	WASF3	KIAA0555	EIF2C1
F	LPXN	ATP1B3	VAMP4		PXN	DCTN1	WASF2		MSN	KIF2C	ITGB7	CAV3
G	PLS3	ADRM1	ARPC1B	TUBG1	MRLC2	GTSE1	TAGLN2	ARPC5	ACTN2	MYO3A	ACTN1	FILIP1
H	ARP3BETA	CNN3	LCP1	VIM	NINJ2	PFN2	PARVA	TPM2	MYO1A	ANLN	FSCN1	KRT8

Table 2: Representative samples of cells transfected with 96 constructs in the construct map

2.3.3 Image Processing

After the high-content imaging, the images are analyzed with the computational framework described in Methods section, involving individual cell segmentation, Intensity transformation of subcellular compartments, protein segmentation based on correlation coefficients, and co-localization quantification of proteins in subcellular components. For cell segmentation, a multi-resolution segmentation technique provided by Definiens Developer was used (Baatz and Schäpe 2000). In the intensity transformation step, the fitting parameters for the exponential function are: $t_m=1.054$, $t'_m=1.53$, $t''_m=2.241$.

A co-localization matrix is generated with the dimension of 89 proteins $\times$ 5 subcellular compartments $\times$ the number of cells transfected with each protein.

2.3.4 Cell Selection

Before clustering the proteins, I applied a cell selection procedure based on nucleus area to delete part of the under-segmented cells, as the nucleus area should have little variance in normal cells. The histogram of Nucleus Area feature is represented as a two-peak curve (Fig.4), one peak is relatively smaller than the other. From biology I know that the higher peak shows the population of the normal nucleus and the smaller peak appears at the position where the area is twice of the normal nucleus area, showing the under-segmented two-connected nuclei. A single Gaussian-fit is applied to the histogram of nucleus area and the interval, which contains 90% of the values from the fitting distribution, was computed. The interval is [302,997]. Although the interval still contains some under-segmented nuclei, it successfully removes many of the miss-segmented cells.

In the current image dataset, the low transfection efficiency largely affects the final results of the co-localization analysis: a large number of cells are not transfected pulling down the total sample number; while the over-expressed ones show abnormal morphology and consequently abnormal co-localization results. After the transfection, the cytoskeletal proteins will initially display their normal subcellular locations or their functional locations. But in the over-expressed situation, as the cytoskeletal proteins localize at low affinity sites within the cell, producing noise and reducing the significance of the functional locations of the particular cytoskeletal protein (Fig. 5). Thus, I performed a GFP-intensity analysis to find the optimized intensity interval to eliminate abnormal-transfected cells.

In the GFP intensity analysis, the transfected cells are clustered into several GFP-intensity groups. By computing the co-localization values for each GFP intensity group, the co-localization trends along with the increasing GFP intensities are investigated and I concluded that the co-localization values do change greatly with the increasing GFP intensity. In the high GFP intensity intervals (greater than 50), I found that the co-localization values in the Nucleus compartment increased, indicating that the over-expressed cells could round up, as dead cells or toxic cells, which been proved by observation. This phenomenon affects the overall co-localization results, especially the constructs with low transfection efficiency. To keep the particular co-localization pattern as well as to avoid noise, the intensity interval [20, 30] seems to be a good choice for the cell selection (Fig. 6).

Figure 4 Cell selection based on Nucleus size. Scatterplot shows the histogram of nucleus size of each segmented cells. Gaussian fit for the first peak is shown in red curve.

Figure 5 Cell samples for different GFP intensity levels. Cell samples are taken from cell set with Cofilin1 (CFL1) protein transfected, and the GFP intensities of the cell samples are from 5 to 60. The cells with GFP levels higher than 50 would be identified as toxic or dead cells.

Figure 6 Cell selection based on GFP intensity

2.3.5 Co-localization Indexing

In order to provide a standard comparison between subcellular compartments, I perform k-means clustering separately on all 5 co-localization values. The sums of squared distances are examined to determine the best number of clusters. For each K, the K-means clustering is replicated 100 times to mitigate the effects of different initial conditions. Three is decided as the number of clusters as most of the co-localization values show inflection on it. The cluster labels from 1 to 3 are assigned to each protein to represent its co-localization degree for a specific subcellular compartment.

2.3.6 Protein Clustering

K-means clustering is performed again based on the cluster label of each protein for further protein classification. The sums of squared distances are examined to determine the best number of clusters. For each K, the K-means clustering is replicated 100 times to mitigate the effects from different initial conditions. Four is chosen as the number of the clusters and all the proteins are clustered into four clusters (Table 3): cluster 1 – 21 proteins with equally distributions within cells; cluster 2 – 21 proteins with high co-localization in plasma membrane; cluster 3 – 34 proteins with high co-localization in actin and cytosol; and cluster 4 – 13 proteins which are toxic to cells leading to cell round up. In Table 2 all proteins in each cluster are listed. As cytoskeletal proteins dynamically function within cells, the co-localization profiles will provide a distribution ratio among the subcellular compartments rather than one specific compartment. Although the exact functions of most of proteins remain unclear, the results can be validated with literature research. It is noticeable that some proteins in the same protein family are clustered

together with similar co-localization profiles, such as TAGLN and TAGLN2, ITGB1 and ITGB2, MYO3A and MYO1A, etc. Another validation is the comparison of reported functions of proteins and co-localization profiles. For example, cluster 3 shows high co-localization in actin compartment, and 22 of 34 proteins in this cluster are reported as having relative functions with the actin network.

Comparing the proteins within the same cluster and in different clusters, proteins with similar co-localization profiles are considered to have similar functions. For example, a set of proteins with unclear function such as filamin A interacting protein 1 (FILIP1) and tropomyosin 1 (TPM1), are seen to have similar profiles showing significantly high co-localization values in plasma membrane (cluster 2) together with other proteins having related functions with plasma membrane such as integrin beta 1 (ITGB1), integrin beta 2 (ITGB2), and villin 2 (VIL2).

Table 3. Co-localization of Protein Clusters

	<i>Nucleus</i>	<i>CytoP</i>	<i>CytoS</i>	<i>Actin</i>	<i>PM</i>	<i>protein No.</i>
Cluster1	52.42%± 4.54%	47.58%± 4.54%	16.26%± 3.31%	49.40%± 5.41%	14.37%± 5.81%	21
Cluster2	56.63%± 8.48%	43.37%± 8.48%	7.85% ±2.66%	43.96%± 4.66%	37.60%± 4.18%	21
Cluster3	34.70%± 6.08%	65.30%± 6.08%	18.62%± 3.36%	53.23%± 7.53%	20.54%± 5.50%	34
Cluster4	75.72%± 9.07%	24.28%± 9.07%	6.81%±3. 07%	45.99%± 13.04%	15.57%± 7.27%	13

Table 4. Proteins in Protein Clusters

	<i>protein No.</i>	<i>Protein Brief Name</i>
Cluster1	21	CORO2B, PDLIM3, DNM2, TAGLN, TNS, TEK3, PTK9, PLS1, JAMIP2, ATP1B3, VAMP4, PXN, MSN, ADRM1, MRLC2, TAGLN2, ARPC5, VIM, NINJ2, PFN2, PARVA
Cluster2	21	TUBA6, TGOLN2, CORO1B, ATP6V1C1, TAGLN3, VIL2, TUBD1, TUBGCP3, PARVG, RDX, CAPZA3, ACTG1, PTK9, TPM1, ITGB1, ITGB2, ADAM15, ARHGEF6, EIF2C1, WASF2, FILIP1
Cluster3	34	DCAMKL1, WASPIP, KLHDC9, CTTN, VIL1, ACTB, WASL, ZYX, CDC42, CFL1, EVL, TUBE1, CLIP3, DSTN, PAK4, VASP, LPXN, KIF2C, ITGB7, PLS3, ARPC1B, TUBG1, GTSE1, ACTN2, MYO3A, ACTN1, ARP11, CNN3, LCP1, TPM2, MYO1A, ANLN, FSCN1, KRT8
Cluster4	13	WAS, ATP5G2, ITGB3BP, FSCN3, GJB2, TUBA1B, KPTN, TEK1, TUBB2, KAT5, WASF3, DCTN1, CAV3

2.4 Conclusion

In this chapter, I developed a computational framework and optimized every step in the framework including cell segmentation, protein segmentation, intensity transformation of subcellular compartments and cell selection to quantify the subcellular localization of cytoskeletal proteins with one specific defined single co-localization measurement. The framework is applied to a two-dimensional image set, containing around 7120 images of cells transfected with 89 cytoskeletal protein constructs. The subcellular localizations of those cytoskeletal proteins are quantified and localization patterns are investigated to provide references in investigation of protein functions. Proteins with unknown functions can be investigated by comparing with co-localization profiles generated in a cytoskeletal protein library.

For image-based subcellular localization quantification, two-dimensional analysis is not sufficient. In future work, the whole framework will be transferred to the three-dimensional domain. The quantification of subcellular localization on three-dimensional space will provide more accurate results. The quantification of subcellular localization will greatly benefit the investigation of functions of cytoskeletal proteins by providing the similarities of subcellular localization patterns of proteins to gain insights of the similarities of functions of proteins.

Chapter 3

Subcellular Localization Prediction with Co-Localization Features and Protein Sequence Features

3.1 Introduction

In Chapter 2, I have discussed the performance of a single co-localization measurement in quantifying and profiling subcellular localization of cytoskeletal proteins. However, besides the Weighted overlapping area, which was defined in Chapter 2, there are often well-defined co-localization measurements as mentioned in Chapter 1.3.2. Those co-localization measurements have the capability to describe different aspects of the co-localization phenomenon with different definitions. For example, Pearson Correlation Coefficient can clearly present the linear relationship between the intensities of pixels, while Manders' overlap coefficient mostly presents the overlapping coefficient between two channels.

In this chapter, selected co-localization measurements will be extracted and collected as a co-localization feature set, which will then be inputted into a developed computational framework for prediction of subcellular localization based on support vector machines (Fig. 7). Reported subcellular localization from protein databases will be used as ground truth in validation steps.

Furthermore, as mentioned in 1.3.1, protein features, other than image features, are shown their power in predicting subcellular localization. In this chapter, I will also describe my attempts in predicting subcellular localizations with co-localization features and protein sequence features, thus to compare their performance in the prediction process.

Figure 7 The Framework of the classifier training process. The whole framework can be separated into two main processes: Images processing and localization classification.

3.2 Feature Sets

3.2.1 Co-localization Features

Here we skip the description of the image processing steps, which have been discussed a lot in previous chapter. After the segmentation of cells and intensity transformation of subcellular localizations, a set of 25 co-localization features are extracted, containing four groups of correlation measurements describing different aspects of the co-localization between protein and particular subcellular compartments (Table 5).

Table 5. Co-localization Features

Feature No.	WOA	ICQ	MOC	MOCp	PCC
Nucleus	1	6	11	12	21
Cytoplasm	2	7	13	14	22
Cytosol	3	8	15	16	23
Actin	4	9	17	18	24
Plasma Membrane	5	10	19	20	25

a. Weighted Overlapping Area

The weighted overlapping area (WOA) is defined to provide quantification about the percentage protein localized in a compartment. The WOA of a protein p in a compartment c is defined as:

$$\text{WOA}(c, p) = \frac{\sum_{x \in W_{p,c}} \tilde{f}_p(x) \cdot \tilde{f}_c(x)}{\sum_{x \in W_{p,c}} \tilde{f}_p(x)} ; \quad (10)$$

where $W_{p,c}$ represents the set of pixels in the region occupied by the protein p correlated with compartment c . $\tilde{f}_p(x)$ is the normalized intensity distribution of the GFP channel highlighting protein p and $\tilde{f}_c(x)$ is the intensity transformation of the compartment c .

b. Pearson's Correlation Coefficient

The Pearson's Correlation Coefficient (*PCC*) is used to describe the linear correlation between protein intensities and intensity transformations of subcellular compartments:

$$PCC(c, p) = \frac{\sum_{x \in W_{p,c}} (\tilde{f}_p(x) - P) \times (\tilde{f}_c(x) - C)}{\sqrt{\left[\sum_{x \in W_{p,c}} (\tilde{f}_p(x) - P)^2 \times \sum_{x \in W_{p,c}} (\tilde{f}_c(x) - C)^2 \right]}}; \quad (11)$$

Where P and C is the average intensities over the whole region occupied by the protein p correlated with compartment c .

c. Manders' Overlap Coefficient

The Manders' Overlap Coefficient (*MOC*) is based on the Pearson correlation coefficient without the average intensity values (9), and an Overlap coefficient for every channel is also defined to describe the correlations separately. In this case, I only looked the overlap coefficient for protein channel (*MOC_p*) (10):

$$MOC(c, p) = \frac{\sum_{x \in W_{p,c}} \tilde{f}_p(x) \times \tilde{f}_c(x)}{\sqrt{\left[\sum_{x \in W_{p,c}} (\tilde{f}_p(x))^2 \times \sum_{x \in W_{p,c}} (\tilde{f}_c(x))^2 \right]}}; \quad (12)$$

$$MOC_p(c) = \frac{\sum_{x \in W_{p,c}} \tilde{f}_p(x) \times \tilde{f}_c(x)}{\sum_{x \in W_{p,c}} (\tilde{f}_p(x))^2}; \quad (13)$$

d. Intensity Correlation Quotient

The Intensity Correlation Quotient (*ICQ*) value is based on the intensity correlation coefficient ρ . The correlation coefficient at a pixel is:

$$\rho(\tilde{f}_p, \tilde{f}_c, x) = (\tilde{f}_p(x) - P)(\tilde{f}_c(x) - C), \quad x \in W_{p,c}; \quad (14)$$

The *ICQ* is defined as the ratio of the positively correlated pixels and the negatively correlated pixels in the region $W_{p,c}$. The correlation coefficient for two denoted areas is:

$$ICQ(\tilde{f}_p, \tilde{f}_c, W_{p,c}) = \frac{\delta^+}{\delta^+ + \delta^-} - 0.5, \quad (15)$$

where $\delta^+ = \sum_{x \in W_{p,c}} \delta(\rho(x) > 0)$ donates the total number of positively correlated pixels, and

$\delta^- = \sum_{x \in W_{p,c}} \delta(\rho(x) < 0)$ donates the total number of negatively correlated pixels. The

range of *ICQ* falls between $[-0.5, 0.5]$. When $ICQ \approx 0$, random correlation; $-0.5 \leq ICQ < 0$, negative correlation; $0 < ICQ \leq 0.5$, positive correlation.

3.2.2 Protein Sequence Features

The sequence feature set consists of (a) 20 Amino Acid Compositions, (b) 400 Dipeptide Composition, (c) 360 Autocorrelation descriptors, (d) 21 Composition, 21 transition, 105 distribution (CTD), (e) 160 Quasi-sequence-order descriptors (f) 80 Amphiphilic Pseudo amino acid composition and (g) 4 Total amino acid property.

A protein or peptide sequence with N amino acid residues is expressed as:

$R_1, R_2, R_3, \dots, R_N$, where R_i represents the residue at the i -th position in the sequence. The labels i and j are used to index amino acid position in a sequence and r and s are used to index the amino acid type.

a. Amino Acid Compositions

The amino acid composition is the fraction of each amino acid type within a protein. The fractions of all 20 natural amino acids are calculated as:

$$f(r) = \frac{N_r}{N} \quad r=1, 2, 3, \dots, 20 \quad (16)$$

Where N_r is the number of the amino acid type r and N is the length of the sequence.

b. Dipeptide Composition

The dipeptide composition gives 400 features, defined as:

$$fr(r, s) = \frac{N_{rs}}{N-1} \quad r, s=1, 2, 3, \dots, 20 \quad (17)$$

where N_{rs} is the number of dipeptide represented by amino acid type r and s .

c. Autocorrelation Descriptors

Autocorrelation descriptors are defined based on the distribution of amino acid properties along the sequence. The amino acid properties used here are various types of amino acids index. Here the indexes of Hydrophobicity index, Relative mutability, Polarity and Hydrophilicity are used. Three type of autocorrelation descriptors are used here and are described as following.

All the amino acid indices are centralized and standardized before the calculation, i.e.

$$P'_r = \frac{P_r - \bar{P}}{\sigma} \quad (18)$$

where $\bar{P}$ is the average of the property of the 20 amino acids:

$$\bar{P} = \frac{\sum_{r=1}^{20} P_r}{20} \quad (19)$$

And

$$\sigma = \sqrt{\frac{1}{20} \sum_{r=1}^{20} (P_r - \bar{P})^2} \quad (20)$$

(1) Normalized Moreau-Broto autocorrelation descriptors

Moreau-Broto autocorrelation descriptors application to protein sequences may be defined as:

$$AC(d) = \sum_{i=1}^{N-d} P_i P_{i+d} \quad d=1, 2, 3, \dots, \text{nlag} \quad (21)$$

Where d is called the lag of the autocorrelation and P_i and P_{i+d} are the properties of the amino acids at position i and i+d , respectively. nlag is the maximum value of the lag.

The normalized Moreau-Broto autocorrelation descriptors are defined as:

$$ATS(d) = \frac{AC(d)}{N-d} \quad d=1, 2, 3, \dots, \text{nlag} \quad (22)$$

(2) Moran autocorrelation

Moran autocorrelation descriptors application to protein sequence may be defined as:

$$I(d) = \frac{\frac{1}{N-d} \sum_{i=1}^{N-d} (P_i - \bar{P})(P_{i+d} - \bar{P})}{\frac{1}{N} \sum_{i=1}^N (P_i - \bar{P})^2} \quad d=1, 2, 3, \dots, 30. \quad (23)$$

$\bar{P}$ is the average of the considered property P along the sequence, i.e.,

$$\bar{P} = \frac{\sum_{i=1}^N P_i}{N} \quad (24)$$

(3) Geary autocorrelation Descriptors

Geary autocorrelation descriptors application to protein sequence may be defined as:

$$C(d) = \frac{\frac{1}{2(N-d)} \sum_{i=1}^{N-d} (P_i - P_{i+d})^2}{\frac{1}{N-1} \sum_{i=1}^N (P_i - \bar{P})^2} \quad d=1, 2, 3, \dots, 30. \quad (25)$$

d. Composition, transition, distribution (CTD)

The amino acids are divided in three classes according to its attribute and each amino acid is encoded by one of the indices 1, 2, 3 according to which class it belonged.

The attributes used here include hydrophobicity, normalized van der Waals volume, polarity, and polarizability, as in the references. The corresponding division is in the table 6.

Table 6 Amino Acid attributes and the Division of the Amino Acids

Property	Class 1	Class 2	Class3
Hydrophobicity	Polar	Neutral	Hydrophobicity
	R,K,E,D,Q,N	G, A, S,T,P,H,Y	C,L,V,I,M,F,W
Normalized Van Der Waals Volume	0-2.78	2.95-4.0	4.03-8.08
	G,A,S,T,P,D	N,V,E,Q,I,L	M,H,K,F,R,Y,W
Polarity	4.9-6.2	8.0-9.2	10.4-13.0
	L,I,F,W,C,M,V,Y	P,A,T,G,S	H,Q,R,K,N,E,D
Polarizability	0-1.08	0.128-0.186	0.219-0.409
	G,A,S,D,T	C,P,N,V,E,Q,I,L	K,M,H,F,R,Y,W
Charge	Positive	Neutral	Negative
	KR	ANCQGHILMFPSTWYV	DE
Secondary Structure	Helix	Strand	Coil
	EALMQKRH	VIYCWFT	GNPSD
Solvent Accessibility	Buried	Exposed	Intermediate
	ALFCGIVW	PKQEND	MPSTHY

(1) Composition

It is the global percent for each encoded class in the sequence. Composition can be defined as:

$$C_r = \frac{n_r}{N} \quad r=1,2,3 \quad (17)$$

Where n_r is the number of r in the encoded sequence and N is the length of the sequence.

(2) Transition

A transition from class 1 to 2 is the percent frequency with which 1 is followed by 2 or 2 is followed by 1 in the encoded sequence. Transition descriptor can be calculated as:

$$T_{rs} = \frac{n_{rs} + n_{sr}}{N - 1} \quad rs = \text{"12"}, \text{"13"}, \text{"23"} \quad (18)$$

Where n_{rs} , n_{sr} is the numbers of dipeptide encoded as "rs" and "sr" respectively in the sequence and N is the length of the sequence.

(3) Distribution

The "distribution" descriptor describes the distribution of each attribute in the sequence. There are five "distribution" descriptors for each attribute and they are the position percent in the whole sequence for the first residue, 25% residues, 50% residues, 75% residues and 100% residues, respectively, for a specified encoded class.

e. Quasi-sequence-order descriptors

The quasi-sequence-order descriptors are proposed by K.C.Chou, et.al. They are derived from the distance matrix between the 20 amino acids.

(1) Sequence-order-coupling numbers

The d_{th} -rank sequence-order-coupling number is defined as:

$$\tau_d = \sum_{i=1}^{N-d} (d_{i,i+d})^2 \quad d=1, 2, \dots, \text{maxlag} \quad (19)$$

Where $d_{i,i+d}$ is the distance between the two amino acids at position i and $i+d$. maxlag is the maximum lag and the length of the protein must be not less than maxlag .

(2) Quasi-sequence-order (QSO) descriptors

For each amino acid type, a quasi-sequence-order descriptor can be defined as:

$$X_r = \frac{f_r}{\sum_{r=1}^{20} f_r + w \sum_{d=1}^{\text{maxlag}} \tau_d} \quad r=1, 2, 3, \dots, 20 \quad (20)$$

Where f_r is the normalized occurrence for amino acid type r and w is a weighting factor ($w=0.1$).

These are the first 20 quasi-sequence-order descriptors. The other 30 quasi-sequence-order are defined as:

$$X_d = \frac{w\tau_{d-20}}{\sum_{r=1}^{20} f_r + w \sum_{d=1}^{\text{maxlag}} \tau_d} \quad d=21, 22, 23, \dots, 20+\text{maxlag} \quad (21)$$

In addition to Schneider-Wrede physicochemical distance matrix used by Chou et al, another chemical distance matrix by Grantham is also used here.

f. Amphiphilic Pseudo amino acid compositions

Let $H_1^0(i)$, $H_2^0(i)$, $M^0(i)$ ($i=1,2,\dots,20$) be the original hydrophobicity values (Jiri Damborsky, 1998), the original hydrophilicity values (Hopp-Woods,1981) and the

original side chain masses of the 20 natural amino acids, respectively. They are converted to following qualities by a standard conversion:

$$H_1(i) = \frac{H_1^0(i) - \sum_{i=1}^{20} \frac{H_1^0(i)}{20}}{\sqrt{\frac{\sum_{i=1}^{20} [H_1^0(i) - \sum_{i=1}^{20} \frac{H_1^0(i)}{20}]^2}{20}}} \quad (22)$$

$$H_2(i) = \frac{H_2^0(i) - \sum_{i=1}^{20} \frac{H_2^0(i)}{20}}{\sqrt{\frac{\sum_{i=1}^{20} [H_2^0(i) - \sum_{i=1}^{20} \frac{H_2^0(i)}{20}]^2}{20}}} \quad (23)$$

$$M(i) = \frac{M^0(i) - \sum_{i=1}^{20} \frac{M^0(i)}{20}}{\sqrt{\frac{\sum_{i=1}^{20} [M^0(i) - \sum_{i=1}^{20} \frac{M^0(i)}{20}]^2}{20}}} \quad (24)$$

The hydrophobicity and hydrophilicity correlation functions are defined respectively as:

$$H_{i,j}^1 = H_1(i)H_1(j) \quad \& \quad H_{i,j}^2 = H_2(i)H_2(j) \quad (25)$$

From these qualities, sequence order factors can be defined as:

$$\begin{aligned}
 \tau_1 &= \frac{1}{N-1} \sum_{i=1}^{N-1} H_{i,j+1}^1 \\
 \tau_2 &= \frac{1}{N-1} \sum_{i=1}^{N-1} H_{i,j+1}^2 \\
 \tau_3 &= \frac{1}{N-2} \sum_{i=1}^{N-2} H_{i,j+2}^2 \\
 \tau_4 &= \frac{1}{N-2} \sum_{i=1}^{N-2} H_{i,j+2}^2 \\
 &\vdots \\
 \tau_{2\lambda-1} &= \frac{1}{N-\lambda} \sum_{i=1}^{N-\lambda} H_{i,j+\lambda}^1 \\
 \tau_{2\lambda} &= \frac{1}{N-\lambda} \sum_{i=1}^{N-\lambda} H_{i,j+\lambda}^2
 \end{aligned} \quad \lambda < N \quad (26)$$

Then a set of descriptors called “Amphiphilic pseudo amino acid composition” (APAAC) are defined as:

$$p_u = \begin{cases} \frac{f_u}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{2\lambda} \tau_j} & (1 \leq u < 20) \\ \frac{w\tau_u}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{2\lambda} \tau_j} & (20+1 \leq u \leq 20+2\lambda) \end{cases} \quad (27)$$

Where w is the weighting factor and is taken as w=0.5 in profeat as in the work of Kuo-Chen Chou (2005).

g. Total amino acid property

The “total amino acid property (TAAP)” descriptor for a property i is defined here as:

$$P_{total}(i) = \sum_{j=1}^N P_j^i \quad (28)$$

Where P_j^i is the property i of amino acid R_j and N is the length of the sequence.

Overall the final feature set consists of 1196 features: 25 co-localization features and 1171 sequence features. (Table 7,8)

Table 7 Co-localization Features

Colocalization Features	Feature No.
Weighted Overlapping Area (WOA)	5
Pearson's Correlation Coefficient (PCC)	5
Manders' Overlap Coefficient (MOC)	10
Intensity Correlation Quotient (ICQ)	5
Subtotal	25

Table 8 Sequence Features

Sequence Features (Max lag=30)	Feature No.
Amino Acid Compositions	20
Dipeptide Compositions	400
Autocorrelation Descriptors	$3 \times 4 \times 30$
Composition, transition, distribution (CTD)	$3 \times 7 + 3 \times 7$ $+ 3 \times 5 \times 7$
Quasi-sequence-order descriptors (QSO)	$[30 + (20 + 30)] \times 2$
Amphiphilic Pseudo amino acid compositions (APAAC)	$20 + 2 \times 30$
Total amino acid property	4
Subtotal	1171

3.3 Localization Classification

The framework of localization classification is illustrated in Fig. 8, and the detailed steps are described in this section.

Figure 8 The framework of localization classification

3.3.1 Ground Truth and Training Set

After co-localization feature extraction, the data can be considered as a co-localization matrix with the dimension of 89 proteins X 25 co-localization features X the number of cells transfected with each protein. We further average the co-localization feature values for all the cells transfected with particular proteins, thus concentrating the data into an 89 X 25 matrix.

In order to validate the subcellular localization of those cytoskeletal proteins, the reported subcellular localizations for each cytoskeletal protein are obtained from the Universal Protein Resource knowledgebase (UNIPROT <http://uniprot.org>). Out of 89 proteins, 73 cytoskeletal proteins have reported subcellular localizations from the database, thus the data of these proteins is used for the training and validation of the classifiers.

3.3.2 Bootstrapping

The feature set of 73 cytoskeletal proteins is sub-sampled into training set (80% proteins) and validation set (20% proteins). This process is repeated 50 times to create variations among the training sets. For a particular training set from 50 sub-samplings, the diversities are generated by the bootstrapping method (Wu 1986) which will help to achieve more robust selection results. By sampling with replacement, a particular training set is 5-fold sub-sampling 50 times to created 50 bootstrap samples and class-specific feature selection is applied to each of the bootstrap sample. 50 out-of-bag (OOB) samples that are left out of the sampling are used later to evaluate the size of the best subset.

3.3.3 Class Binarization

Those training and validation sets are used to find the best feature set from the 25 features, and thus build classifiers for different subcellular compartments. As mentioned previously, the dynamic characteristic of the cytoskeletal proteins, may have several localizations instead of a fixed localization, thus we decided to build classifiers for every single subcellular compartment, expecting the combined classification results from different classifiers will better describe the dynamic localizations of cytoskeletal proteins. All the 73 proteins with reported subcellular localizations are located in the cytoplasm; none of them are located in Nucleus only. Therefore we only trained 4 classifiers for subcellular compartments: Nucleus, Cytosol, Actin, and Plasma membrane. The four-class classification problem is then transformed into four binary classification problems with a one-against-all class binarization approach. Each bootstrap sample is re-labelled accordingly for each class, where the instances of the certain class are assigned 1 while the instances of all the other classes are assigned 0.

3.3.4 Class Balancing

For the binary classification of each class, we further balance the two classes by sampling with replacement on the class with fewer samples. This procedure is necessary since class binarization in the previous step causes significant imbalance in the two classes which will affect the feature selection in the next steps (Pineda-Bautista and Carrasco-Ochoa 2011).

3.3.5 Feature Ranking

Feature ranking and selection is an important procedure before we train the certain classifier with the extracted feature sets, by eliminating unimportant features and selecting the best subset which contribute most to the accuracy of classification. Support Vector Machine (SVM) (Cortes and Vapnik 1995) is used as the classifier in this study, which is proven to offer state-of-the-art classification performance in different applications, especially for the small sample-to-feature ratio cases (Guyon, Weston et al. 2002). We use a support vector machine recursive feature elimination (SVM-RFE) approach (Guyon, Weston et al. 2002) to rank the features, as the method makes use of the same classification model as the classifier we choose. SVM-RFE was introduced as a multivariate feature ranking method that uses SVM weights as the ranking criterion of features. Starting from using all the features, a SVM is trained and features are sorted by their weight (importance) in the classification. The feature with the least weight is least important and is eliminated from the feature set and a SVM is trained again using the new feature set after elimination, until all the features have been removed. All the features are hence ranked during this recursive elimination procedure.

As we aim to classify the proteins into one or more of the four classes (subcellular compartments), there are two ways to perform feature selection for multi-class classification problem. We can either select the features for all the classes or for each class separately. The former method selects a single subset of features to discriminate all the classes while the class-specific method selects a subset for each class and the subsets could be different between different classes (Pineda-Bautista and Carrasco-Ochoa 2011). The class-specific method takes advantage of the notion that there may be features which

are important for a specific class but not included in the best subset for all the classes. In the present study, we use the SVM-RFE method as the feature selector in a class-specific feature selection framework.

SVM-RFE method is applied to each bootstrap sample for the binary classification of each class and let $F_{i,j,n}$ denotes the rank of feature n in bootstrap j for class i , where the best feature is ranked 1 and the worst feature is ranked N , which is the number of the features. The linear SVM sensitivity parameter is estimated using the same approach by Mundra and Rajapakse (Mundra and Rajapakse 2010). We sum up the ranks for each feature over all the bootstrap samples from 1 to J for class i with weighted linear aggregation approach and the ensemble rank of feature n for class i is calculated as

$$F_{i,n} = \sum_{j=1}^J w_{i,j} F_{i,j,n},$$

which is similar to what has been proposed in (Abeel, Helleputte et al.

2010). We train a linear SVM on each bootstrap sample separately for each class and the classifier is evaluated on the out-of-bag samples based on the AUROC measurement (Hanley 1982), which is assigned to $w_{i,j}$ as the weight of bootstrap sample j for class i .

The feature with the smallest ranking in $F_{i,n}$ is assigned 1 as the best feature and the feature with the biggest ranking in $F_{i,n}$ is assigned N as the worst.

3.3.6 Selection of the Best Feature Subset

For each class, different numbers of the features are selected according to the ensemble rankings from the top order to train a linear SVM by bootstrap samples and to be evaluated on the out-of-bag samples. The subset of the features which lead to the highest

average accuracy over all out-of-bag samples is finally selected. The selected subsets of features for each class are further merged together as $F = \cup_{i=1}^4 F_i$, where F_i denotes the selected subset of class i .

3.3.7 Classifier Training and Evaluation

A one-against-all SVM is trained on each particular training set from 50 sub-samplings with the selected features by the method elaborated in previous section. The classifier is then evaluated on the particular test set. We measure the classification performance by the average classification accuracy over all test sets from 50 sub-samplings. The accuracy is the percentage of test sets that are correctly classified. A better feature set tends to generate higher accuracy.

All the features from different feature sets are put together to perform feature selection. The proportion of the selected features from a certain feature set over all the selected features could change comparing to the original proportion of the features from this certain feature set over all the features before feature selection. The increased proportion could be an indicator to demonstrate that this certain feature set is relatively more important than others.

3.4 Experiments and Discussion

3.4.1 Subcellular Localization Prediction with Co-Localization Features

1. Feature Selection and Classifier Accuracy

The method described above is applied on the cytoskeletal protein dataset, and four classifiers are generated for four subcellular compartments: Nucleus, Actin, Plasma Membrane, and Cytosol. The performance of these classifiers is examined (Fig. 9). The accuracies of Nucleus, Actin and Cytosol classifiers are all above 0.6, an acceptable result; while the Plasma Membrane classifier is above 0.55 (Fig. 9B). The reason for the low accuracies might be the small dataset. For example, there are only a few proteins from the training set have reported localization with the plasma membrane, thus, during the class balancing step, data for those proteins will be duplicated a lot to balance the sample size, but the information is just repeated. Another possible reason could be the “unsure” ground truth. The subcellular localization obtained from the Uniprot database would be correct; but the cytoskeletal proteins wander a lot inside the cells and most of them have unsure functions. They may locate in the subcellular compartments without reporting on the Uniprot database. During the validation process, we assume that the protein doesn’t locate in the compartments if it is not mentioned in the Uniprot database. But actually, no reporting does not mean there is no localization.

Figure 9. Performance of the localization classifiers: (A) the percentage of features selected for each classifier. (B) The accuracy of each classifier.

The nucleus classifier has a significantly larger feature subset compared to the other three classifiers (Fig. 9A). We examined the feature frequency of four classifiers (Fig. 10). The feature frequency of the Nucleus classifier distributes almost uniformly (Fig. 10A), suggesting those features have similar possibility of being chosen, and the Nucleus classifier doesn't rely on several important features, but a combination of many features with similar importance for the classification. On the other hand, several features showed their importance with high selection frequency in the Actin, Plasma Membrane, and Cytosol classifiers (Fig. 10B-D). For example, surprisingly, feature No.18, the MOCp on the Actin compartment have a frequency greater than 0.8 in all three classifiers, especially in Plasma Membrane classifier.

Figure 10. The feature frequency of four classifiers during the 50 times of feature selection with random sub-sampling training sets. (A)Nucleus Classifier (B) Actin Classifier (C) Plasma Membrane Classifier (D) Cytosol Classifier

Figure 11. The proportion changes of features grouped with different measure methods: (A) weighted overlapping area, (B) intensity correlation quotient, (C) overlap coefficient, and (D) Pearson's correlation coefficient.

Therefore, we further examined proportion changes of different feature groups before and after the feature selection. For co-localization features grouped with different measurements (Fig. 11), WOA features remains unchanged in the Actin classifier, while

they decreased in the other three classifiers, especially in Plasma Membrane (Fig. 11A), indicating the WOA features are not distinguishable for the localization classification as other three types of features. ICQ features increased a lot in Actin and Cytosol classifiers (fig. 11B), but decreased in Nucleus and Plasma Membrane classifiers suggesting ICQ features play an important role in the classification of Actin and Cytosol. Similarly, the MOC features are mostly selected with significantly increased proportion changes in the Plasma Membrane classifier but decreased in the Nucleus classifier (Fig. 11C), implying that the MOC features are more crucial in the Plasma Membrane classifier than the other three types of features. And the PCC features clearly showed their importance in the Nucleus classifier, while less importance in other three classifiers (Fig. 11D). To conclude from the view of the classifiers, the Nucleus classifier needs more supporting from PCC features, while the Plasma Membrane classifier chooses more MOC features, and the Actin and Cytosol classifiers rely more on the ICQ features. The WOA features do not have very much proportion change during the feature selection suggesting they have no special contribution to all four classifiers, but they still remain a proportion in the feature subset. Thus we can conclude that the final feature set consist of features from different measure methods and those measurements do describe different aspects of the co-localization and cannot be removed or replaced with each other.

Figure 12. The proportion changes of features grouped with different subcellular compartments: (A) Nucleus classifier, (B) Actin classifier, (C) Plasma Membrane classifier, and (D) Cytosol classifier.

For co-localization features grouped with different subcellular compartments (Fig. 12), features related with the actin compartment are significantly increased in proportion in all four classifiers, while features related with the nucleus and plasma membrane decreased

in all; features related with cytoplasm only increased in the Plasma Membrane classifier, while features related with cytosol only increased in the Cytosol classifier. To conclude from the view of classifiers, the Nucleus classifier mostly relies on actin features, and also a small proportion of features from the other four compartments (Fig. 12A); the Actin Classifier consists of a large proportion of actin features, and equally small proportions of cytoplasm and cytosol features, barely no nucleus and plasma membrane features (Fig. 9B); the Plasma Membrane classifier again contains a large proportion of actin features, and a small proportion of cytoplasm features (Fig. 12C); and the Cytosol classifier has similar high proportions of actin and cytosol features, and a small complement of cytoplasm features (Fig. 12D). Actin is the most important cytoskeletal protein; it is the subunit of microfilaments -- one major component of the cytoskeleton. The Actin network gives mechanical support to cell structures, supports cell motility, and provides trafficking routes through the cytoplasm to support signal transduction. Among the 89 cytoskeletal proteins, large percentages are Actin-binded or functions related to Actin. That could explain why Actin related features occupied a large proportion in the selected feature subset.

2. Localization Prediction

With trained classifiers, we next examined subcellular localization prediction. There are 73 proteins out of 89 have reported localization in the Uniprot. The data on those 73 cytoskeletal proteins are used as training and validation sets. And there are 16 proteins that have no reported localization in the Uniprot. They are classified with the trained classifiers and generate the localization profiles indicating the possible localization they might belong to. The possible functions of those proteins can be looked up in the

GeneCards encyclopedia database (<http://www.genecards.org>). Among the 16 proteins, only 13 proteins have reported functions in the GeneCards database. The predicted localizations are compared with the possible functions to check the prediction accuracy (Table 9).

Table 9. Comparison of Predicted Localization and Possible Functions

Protein Brief	Actin	Nuc	CytoP	PM	CytoS	Reported Function Description
Agreement						
TAGLN	1	0	1	0	0	<u>Actin cross-linking/gelling protein</u> (By similarity). Involved in calcium interactions and contractile properties of the cell that may contribute to replicative senescence
WAS	1	0	1	0	0	<u>Regulates actin filament reorganization</u> via its interaction with the Arp2/3 complex. Important for efficient <u>actin polymerization</u> .
DSTN	1	0	1	0	0	actin depolymerizing protein that severs actin filaments (F-actin) and <u>binds to actin monomers</u> (G-actin)
ATP6V1C1	0	0	1	0	1	Subunit C is necessary for the assembly of the catalytic sector of the enzyme and is likely to have a specific function in its catalytic activity. V-ATPase is responsible for acidifying a variety of intracellular compartments in eukaryotic cells
ARHGEF6	0	0	1	0	1	The encoded protein belongs to a family of <u>cytoplasmic proteins</u> that activate the Ras-like family of Rho proteins by exchanging bound GDP for GTP. Acts as a RAC1 guanine nucleotide exchange factor (GEF)x
MRLC2	1	0	1	0	0	<u>Myosin regulatory subunit</u>
MYO1A	1	0	1	0	1	Molecular motors interact with <u>actin filaments</u> Involved in directing the movement of organelles along <u>actin filaments</u>
CNN3	1	0	1	0	0	Associated with the cytoskeleton. Capable of <u>binding to actin</u> , calmodulin, troponin C, and tropomyosin.
CTTN	1	0	1	0	0	Located in the <u>cytoplasm</u> of cells. Contributes to the <u>organization of the actin cytoskeleton</u> and cell structure. Promoting polymerization and rearrangement of the <u>actin cytoskeleton</u> .
CAPZA3	0	0	1	0	1	<u>actin capping protein</u>
Disagreement						
FILIP1	0	0	1	0	1	acting through a filamin-A/ <u>F-actin axis</u> . May be able to induce the degradation of filamin-A
DCAMKL1	1	0	1	0	0	Microtubule-associated protein. microtubule polymerizing activity. integral component of <u>plasma membrane</u>
Unknown						
TAGLN2	1	0	1	0	0	One of the earliest markers of differentiated smooth muscle. Function unsure

The predicted localization of these 13 cytoskeletal proteins with the trained classifiers are denoted as 1/0 indicating localized/not localized with the subcellular compartments, and the reported possible functions from the GeneCards database are cited and concentrated to short sentences describing functions related to the protein localization with keywords highlighted. All 13 cytoskeletal proteins are divided into three groups: (A) Agreement: predicted localizations match the possible functions, (B) Disagreement: predicted localizations failed to match the possible functions, and (C) Unknown: localizations failed to be concluded from the possible functions. There are nine proteins in group A, and two proteins in group B and C separately. Thus the prediction accuracy on these proteins is $(\text{Agreement})/(\text{Agreement}+\text{Disagreement})=10/(10+1)=90.9\%$. The predicted localization of the nine proteins in group A can be perfectly explained with the reported possible functions, yielding an exact localization prediction. The two proteins in group B state the possible functions associated with the actin compartment or plasma membrane, however, the prediction of those proteins from our classifiers are disagree with their reported functions. And there are one protein TAGLN2 in group C with reported possible functions, but the localizations cannot be concluded from the functions. But note, TAGLN 2 protein is highly homologous with TAGLN in group A, and the predicted localizations of these two proteins are exactly same. Although we cannot conclude that two proteins should have same function, but that could be noted as a reference when investigating the unknown functions of TAGLN2, similarly, to other cytoskeletal proteins with unknown functions.

3.4.2 Subcellular Localization Prediction with Co-Localization Features and sequence features

After both co-localization and sequence feature extraction, the data can be concluded as a feature matrix with the dimension of 89 proteins X (25 co-localization features+1171 sequence features). In this section, I tested two approaches with different flow paths.

1. Feature merging

In this flow path (Fig. 13), we import the sequence features, co-localization features, and combined features separately. Support Vector Machine-Recursive Feature Elimination (SVM-RFE) is employed for the feature ranking. SVM-RFE is one of the most effective feature selection methods, which eliminates redundant features and yields better and more compact feature subset. The four-class classification problem we faced is transformed into six binary classification problems with a one against one class binarization approach. With the selected feature subsets, six binary SVM classifiers are trained. And the outputs from those six classifiers pass through a voting system and thus fuse the final classification decision.

Previously bootstrapping method was used to achieve more robust selection results, which is much more time consuming. In this preliminary test we only choose different top percentage features from the feature ranks to examine the performance. The prediction quality is examined by the jackknife test. The jackknife test is deemed most objective and rigorous compared to sub-sampling test and independent dataset test. The accuracy of four classes from feature merging method is shown in table 10. It is clearly that even we only choose the top percentage features based on the feature ranking, the

classifier from combined features is given equal or higher accuracy compared to only co-localization features or only sequence features.

Figure 13 Flow Path for the Feature merging approach. In this approach, the sequence features and image features are merged first before the following training steps.

Table 10 Accuracy of four classes from feature merging method

<u>Actin</u>	Image Features	Sequence Features	Combine Features
No Selection	63.01%	57.53%	60.27%
Top 1%	63.01%	65.75%	65.75%
Top 5%	63.01%	67.12%	67.12%
Top 10%	56.16%	61.64%	64.38%
Top 25%	63.01%	65.75%	67.12%
Top 50%	60.27%	61.64%	61.64%

<u>Nucleus</u>	Image Features	Sequence Features	Combine Features
No Selection	65.75%	82.19%	80.82%
Top 1%	63.01%	78.08%	75.34%
Top 5%	63.01%	76.71%	82.19%
Top 10%	63.01%	82.19%	86.30%
Top 25%	61.64%	82.19%	83.56%
Top 50%	67.12%	83.56%	80.82%

<u>PM</u>	Image Features	Sequence Features	Combine Features
No Selection	67.12%	78.08%	76.71%
Top 1%	57.53%	69.86%	72.60%
Top 5%	57.53%	75.34%	75.34%
Top 10%	54.79%	76.71%	76.71%
Top 25%	56.16%	80.82%	80.82%
Top 50%	58.90%	79.45%	78.08%

<u>Cytosol</u>	Image Features	Sequence Features	Combine Features
No Selection	60.27%	56.16%	61.64%
Top 1%	64.38%	60.27%	61.64%
Top 5%	64.38%	58.90%	73.97%
Top 10%	69.86%	63.01%	68.49%
Top 25%	61.64%	61.64%	60.27%
Top 50%	61.64%	60.27%	61.64%

2. Vote Merging

In the vote merging flow path, the sequence features and co-localization features are imported into the procedure of feature selection, class binarization, SVM training separately, and the voting will be merged to generate the final decision. (Fig. 14)

Figure 14 Flow Path for the Vote merging approach. In this approach, the SVMs were trained separately with sequence features and image features. And the final prediction result is merged from two SVM flow.

And the accuracy of four classes from vote merging method is shown in Table 11. In vote merging flow path, we can have different feature selections for co-localization features and sequence features, thus we may have different feature combinations as shown in Table 11. The accuracy of Actin and CytoS classes are increased compared to the feature merging method, while Nuc and PM remain the same.

Table 11 Accuracy of Actin and CytoS classes from Vote merging method

Actin		Sequence Features					
		No Selection	Top 1%	Top 5%	Top 10%	Top 25%	Top 50%
Image Features	No Selection	67.12%	60.27%	64.38%	65.75%	67.12%	67.12%
	Top 1%	67.12%	63.01%	69.86%	68.49%	67.12%	64.38%
	Top 5%	67.12%	63.01%	69.86%	68.49%	67.12%	64.38%
	Top 10%	60.27%	58.90%	61.64%	58.90%	58.90%	57.53%
	Top 25%	61.64%	63.01%	64.38%	63.01%	63.01%	64.38%
	Top 50%	64.38%	58.90%	63.01%	64.38%	65.75%	67.12%

CytoS		Sequence Features					
		No Selection	Top 1%	Top 5%	Top 10%	Top 25%	Top 50%
Image Features	No Selection	64.38%	64.38%	65.75%	68.49%	68.49%	65.75%
	Top 1%	65.75%	69.86%	65.75%	65.75%	67.12%	65.75%
	Top 5%	65.75%	69.86%	65.75%	65.75%	67.12%	65.75%
	Top 10%	67.12%	71.23%	69.86%	75.34%	72.60%	68.49%
	Top 25%	67.12%	68.49%	68.49%	71.23%	69.86%	68.49%
	Top 50%	67.12%	67.12%	67.12%	69.86%	71.23%	67.12%

3.5 Conclusion

In this chapter, I developed a computational framework to predict the subcellular localization of cytoskeletal proteins with a series of classifiers trained with co-localization features and protein sequence features extracted by high-content image processing. The method is applied on a two-dimensional image set, containing around 8000 images of cells transfected with 89 cytoskeletal protein constructs. The subcellular localizations of those cytoskeletal proteins are predicted and validated by comparison with the reported functions, which provide references in investigation of protein functions. Co-localization features have shown their capability in subcellular localization prediction and would be a useful complement with other feature set to improve the prediction accuracy.

The poor transfection efficiency greatly limited the accuracy of the classifiers, since the number of transfected cell samples is greatly reduced compared to total cell number. Another limitation is the number of protein samples is not sufficient with less localization variability, thus it is hard to achieve higher classification accuracy due to the limited training and validation set.

For image-based subcellular localization quantification, two-dimensional analysis is not sufficient. In future work, the method will be transferred to the three-dimensional domain. The prediction of subcellular localization on three-dimensional space with higher magnification will provide more accurate results. The prediction of subcellular localization will greatly benefit the investigation of functions of cytoskeletal proteins.

Chapter 4

Co-Localization from 3D Images

4.1 Introduction

In chapters 2 and 3, I have discussed the applications of co-localization measurements in the quantification and prediction of subcellular localization. In this chapter I would like to focus on the co-localization quantification itself. In section 1.1.3 I have introduced the normal process of quantifying co-localization and the most popular co-localization measurements. However, the many different co-localization measurements raises the following questions: How should I choose a co-localization measurement? How should I compare between different co-localization measurements? How should I interpret the meaning of the co-localization based on the calculated values? Before we can answer those questions, we need to understand the differences of these measurements and the performance of these measurements under different conditions.

Quantification of co-localization can be considered as the combination of both co-occurrence and intensity correlation. Thus in this chapter, to address the interpretation problem, I tested and compared the performances of the set of co-localization measurements on synthetic images with variations of co-occurrence and intensity correlation. As a result, I propose a method to separate the aspects of co-occurrence and correlation from the co-localization quantification that would greatly aid in understanding biological structure and function.

Furthermore, I apply my methods to quantify co-localization of transcription factors in

mouse embryonic stem cells from high-resolution 3-dimensional images. The localization of transcription factors on genome is critical for understanding their functions. And therefore the co-localization of those transcription factors would be important in investigating the interaction and relationship between each factor pairs.

4.2 Performance of Co-Localization Measurements

In this part, I describe the performance analysis of the different co-localization measurements on synthetic image sets.

4.2.1 Synthetic Images

To test the performance of the co-localization measurements, two 256*256 8-bit images was generated with Matlab Image Processing Toolbox. Pixel sized objects were randomly distributed on the two images with intensities following Gaussian distributions (mean value of 128 and standard deviation of 128) (Singan, Jones et al. 2011).

1. Co-Occurrence

The co-occurrence was defined as the percentage of pixels of two channels co-occupying the same location (Fig. 15).

The variation of co-occurrence was generated by copying different portions of channel A into channel B. The co-occurrence ranged from 10% to 100% and the positions of portions are chosen randomly. (Adler and Parmryd 2010) (Fig. 16).

Figure 15 The illustration of co-occurrence. In the overlay of red and green channel, there are both pixels occupied same and not same locations. Only the yellow rectangle in the middle of the image is 100% co-occurrence.

Figure 16 A representative sample of Synthetic Image with 30% co-occurrence and 0.6 Correlation. $Img2'$ is the synthetic image generated from $Img2$ to have 30% co-occurrence and 0.6 correlation with $Img1$.

2. Intensity Correlation

The intensity correlation is defined as the degree of the linear relationship between the intensity values of two channels (Fig. 17).

Figure 17 The illustration of Intensity Correlation. Given red channel, three green channel images are shown to demonstrate positive correlation, negative correlation and random correlation. Note for intensity correlation only the intensities of the central rectangle count, as other pixels are not even co-occupied the same location with red channel.

For the aspect of intensity correlation, I have tested two methods to generate the variations of intensity correlation in the synthetic images.

1. Manipulation on intensities

The intensity correlation of the copied portions is manipulated by tuning C_f from 0 to 1 in the equation 29. (Adler and Parmryd 2010).

$$A_{new} = A_{avg} + (A_i - A_{avg}) (1 - C_f) + (B_i - B_{avg})(C_f) \quad (29)$$

Here C_f controls the degree of correlation between two channels. A_i and B_i are intensities of pixel i in channel A and B. A_{avg} and B_{avg} are the averages of channel A and B, respectively. (Fig. 18)

Figure 18 Illustration of Synthetic Images with variations of Correlation. The synthetic images are generated from Img2 to have correlation variations from 0.1 to 1.0 with Img1.

This method was proposed and applied by Adler and Pamryd (Adler and Parmryd 2010) to generate the variation of intensity correlation, however, this formula only generates the overlapping area by copying a certain percentage of intensities onto the synthetic image. Different percentages only represent the change of intensity level, but in the overlapping area the pixels are 100% positive correlated no matter what percentage copied. Thus we proposed a more interpretable way to generate the synthetic images, by manipulating the pixel objects.

2. Manipulation on pixel objects

As co-occurrence has been defined previously, here I only consider the objects in the overlapping area. For pixel-based objects, I could cluster them into three categories: positive correlated, negative correlated and random correlated.

Thus I proposed to manipulate the numbers of pixel-based objects of three categories to generate the variation of intensity correlation. The degree of intensity correlation will indicate the percentage of positive correlated pixels among all overlapped pixels. In the extreme case, here we test positive correlated mixed with negative correlated and positive correlated mixed with random correlated pixels. For example, the synthetic image with 30% Co-occurrence 0.1 correlation means there will be 70% non-overlap pixels and 30% overlap pixels within the image, and within the 30% overlapped pixels there will be $30\% \times 0.1 = 3\%$ perfect correlated and $30\% \times (1 - 0.1) = 27\%$ random/negative correlated in intensity.

4.2.2 Co-localization measurements

In this performance analysis, five global co-localization measurements were analyzed:

1) Pearson's Correlation Coefficient

The Pearson's Correlation Coefficient (PCC) is used to describe the linear correlation between protein intensities and intensity transformations of subcellular compartments:

$$PCC(c, p) = \frac{\sum_{x \in W_{p,c}} (\tilde{f}_p(x) - P) \times (\tilde{f}_c(x) - C)}{\sqrt{\sum_{x \in W_{p,c}} (\tilde{f}_p(x) - P)^2 \times \sum_{x \in W_{p,c}} (\tilde{f}_c(x) - C)^2}}; \quad (30)$$

Where P and C is the average intensities over the whole region occupied by the protein p correlated with compartment c . The value of Pearson's correlation coefficient falls between $[-1, 1]$. $PCC=1$ means perfectly positive correlation, $PCC=-1$ means perfectly negative correlation, and $PCC=0$ means randomly correlation.

Besides calculating the Pearson's correlation coefficient of the whole image, I will also calculate the Pearson's correlation coefficient of the overlapped pixels to check how they correlated with each other. It will be denoted as overlap PCC.

2) Spearman's rank correlation coefficient

The Spearman's rank correlation coefficient (SRC) is defined as the Pearson correlation coefficient (eq. 30) between the ranked variables instead of absolute values.

3) Intensity Correlation Quotient

The Intensity Correlation Quotient (ICQ) value is based on the intensity correlation coefficient ρ . The correlation coefficient at a pixel is:

$$\rho(\tilde{f}_p, \tilde{f}_c, x) = (\tilde{f}_p(x) - P)(\tilde{f}_c(x) - C), \quad x \in W_{p,c}; \quad (31)$$

The ICQ is defined as the ratio of the positively correlated pixels and the negatively correlated pixels in the region $W_{p,c}$. The correlation coefficient for two denoted areas is:

$$ICQ(\tilde{f}_p, \tilde{f}_c, W_{p,c}) = \frac{\delta^+}{\delta^+ + \delta^-} - 0.5, \quad (32)$$

where $\delta^+ = \sum_{x \in W_{p,c}} \delta(\rho(x) > 0)$ donates the total number of positively correlated pixels, and

$\delta^- = \sum_{x \in W_{p,c}} \delta(\rho(x) < 0)$ donates the total number of negatively correlated pixels. The range

of ICQ falls between $[-0.5, 0.5]$. When $ICQ \approx 0$, random correlation; $-0.5 \leq ICQ < 0$, negative correlation; $0 < ICQ \leq 0.5$, positive correlation.

4) Manders' Overlap Coefficient

The Manders' Overlap Coefficient (MOC) is based on the Pearson correlation coefficient without the average intensity values (eq. 30):

$$MOC(c, p) = \frac{\sum_{x \in W_{p,c}} \tilde{f}_p(x) \times \tilde{f}_c(x)}{\sqrt{\left[\sum_{x \in W_{p,c}} (\tilde{f}_p(x))^2 \times \sum_{x \in W_{p,c}} (\tilde{f}_c(x))^2 \right]}}; \quad (33)$$

5) Weighted Overlapping Area

The weighted overlapping area (WOA) is defined to provide quantification about the percentage protein localized in a compartment. The WOA of a protein p in a compartment c is defined as:

$$WOA(c, p) = \frac{\sum_{x \in W_{p,c}} \tilde{f}_p(x) \cdot \tilde{f}_c(x)}{\sum_{x \in W_{p,c}} \tilde{f}_p(x)}; \quad (34)$$

where $W_{p,c}$ represents the set of pixels in the region occupied by the protein p correlated with compartment c . $\tilde{f}_p(x)$ is the normalized intensity distribution of the GFP channel highlighting protein p and $\tilde{f}_c(x)$ is the intensity transformation of the compartment c .

4.2.3 Performance

Different co-localization measurements were calculated on synthetic images: (Fig. 19-21)

Figure 19 Performance of co-localization measurements on synthetic images generated by manipulating intensities.

In the case of calculating from synthetic images generated by manipulating intensities (Fig. 19), Pearson correlation coefficient (PCC) shows non-linear patterns with increased co-occurrence. And the distances between co-occurrence lines are not distributed equally. PCC was shown more sensitivity to the correlation than co-occurrence. Similar to PCC,

both Spearman's rank correlation coefficient (SRC) and Intensity Correlation Quotient (ICQ) provide non-linear pattern and less sensitive to the correlation variations. And also the distances between co-occurrence lines are not distributed equally. The sudden changes of the SRC curves are due to the changes in the number of objects of the synthetic image. Manders' Overlap Coefficient (MOC) provides more linearly pattern than PCC, the distances between co-occurrence lines are more equally distributed, and it show more sensitivity to co-occurrence than correlation, opposite to PCC. Weighted Overlapping Area provides a good pattern, only sensitive to the co-occurrence variations, which could help us determine the co-occurrence of the images. And also the distances between co-occurrence lines are distributed equally. PCC in overlapping area keeps at 1 all the time, suggests that the pixels in overlapping area are always 100% positive correlated.

We then calculated co-localization measurements from synthetic images generated by manipulating numbers of different objects. (Fig. 20-21)

The pattern of those co-localization measurements looks much more smooth than the previous plots, suggesting that the way of generating synthetic images is more similar to how the correlation measurements are defined. And their patterns looks similar in the two synthetic images generated with two extreme situations, suggesting that all those measurements are less sensitive to recognize the populations of pixels correlated or not. They focus on the co-occurrence mostly. In the natural case, those pixels with non-positive correlation do mixed together and it is really hard to distinguish the negative and random correlated populations. WoA measurements still can be used as the indicator for the co-occurrence, and PCC calculated only in the overlapping area could be used to

describe the linear correlation of overlap pixels. Thus, we could separate the co-occurrence and correlation from the co-localization event and analyze them separately. These two parameters yield different aspects of co-localization, which would be helpful in interpreting the biological meaning behind the co-localization phenomenon.

Figure 20 Performance of co-localization measurements on synthetic images generated by mixing positive correlated pixels with random correlated in overlapping area.

Figure 21 Performance of co-localization measurements on synthetic images generated by mixing positive correlated pixels with negative correlated in overlapping area.

4.3 Co-localization of Transcription Factors

After the performance analysis of the co-localization measurements on synthetic images, a new way to describe the co-localization event is purposed as to analyze the co-occurrence and correlation separately, and combined to form an interpretable analysis results.

In this part, I would like to present my attempt in applying this co-localization analysis of transcription factors in mouse Embryonic Stem cells from high-resolution 3-deminational images.

4.3.1 Transcription Factors

1. Transcription factor structure and function, binding to DNA and promoters

Transcription factors recognize and bind to specific DNA sequence to control the transcription of genes by promoting or repressing the recruitment of RNA polymerase. Some transcription factors act on its own while others associate with co-regulators to regulate transcription. There are two types of transcription factors, constitutively active factors and regulatory transcription factors (Brivanlou and Darnell 2002). The constitutively active factors are present in all cells at all times. They regulate genes that are always transcribed such as structural proteins and metabolic enzymes like GADPH. Regulatory transcription factors can be further grouped into developmental transcription factors and signal-dependent transcription factors. Developmental transcription factors are often cell-specific. Their regulation changes during development process. Signal-dependent transcription factors are mostly restricted and present in all cells. They could be initiated or activated once cell exposes to certain extracellular signal.

A typical transcription factor consists of a DNA-binding domain (DBD) and a transactivation domain (TAD) (Latchman 1997). The DNA-binding domain recognizes and binds to specific sequences of DNA (a promoter or enhancer). The transactivation domain interacts with the transcriptional complex (RNA polymerase and general transcription factors) either directly or indirectly through some co-regulators (co-activator or co-repressor). Some transcription factors have signal-sensing domains (SSD), which bind ligands and transmit regulation signals to the rest of the transcription complex.

The structure of the DNA-binding domain of transcription factors can be grouped into several different families, and transcription factors are often classified according to the tertiary structure of their DNA-binding domains (Pabo and Sauer 1992). Some major families of DNA-binding domain structure are:

1) Helix-turn-helix (HTH)

Helix-turn-helix binding domain was the first discovered DNA-binding domain of transcription factors (Pabo and Lewis 1982). It has been well studied and many crystal structures of such domain are available from different proteins (Wintjens and Rooman 1996). The simplest form of helix-turn-helix domain contains two α -helices and a turn in between. The two helices are nearly perpendicular and the turn consists of 3 amino acids (Brennan and Matthews 1989). The second helix from the N-terminal is the “recognition” helix, which aligns with the main groove of DNA. The side chain of the second helix binds specifically to the base sequence on DNA (Harrison and Aggarwal 1990). Besides the simplest structure, there are variations of HTH motif in the turn between the two helices. Longer turns with more than 3 residues (Finney 1990) and shorter turns with 1

residue (Ogata, Hojo et al. 1992) were found. Some other variations of HTH family proteins were found to have β -strands involved in DNA-binding (Brennan 1993).

2) Homeodomain

Homeodomain motif is discovered in a large family of eukaryotic regulatory proteins (Pabo and Sauer 1992) and it is highly conserved during evolution even from prokaryotic DNA-binding proteins (Gehring, Qian et al. 1994). A homeodomain motif consists of 60 residues and contains a HTH motif (Gehring, Qian et al. 1994). The difference between a homeodomain motif and an isolated HTH motif is that a homeodomain motif can form a stable structure itself without interacting with DNA (Sauer, Smith et al. 1988, Affolter, Percival-Smith et al. 1990). A classic homeodomain motif contains 3 α -helices. Helix 1 and helix 2 are antiparallel and packed with each other. Helix 3 is nearly perpendicular to helix 1 and 2, and its hydrophobic face packs with the first 2 helices to form the interior of the protein. Helix 3 makes the most contact to DNA by fitting into the major groove. This positioning also allows the C-terminal residues to make contact with DNA backbone (Pabo and Sauer 1992).

3) Zinc Finger

The zinc finger structure was first discovered in the *Xenopus* transcription factor IIIA (TFIIIA) (Miller, McLachlan et al. 1985). It is one of the major protein-DNA-binding motifs in eukaryotic cells (Pavletich and Pabo 1991). Zinc finger proteins often contain tandem repeats of zinc finger motif. Each zinc finger motif has the sequence pattern of Cys-X₂₋₄-Cys-X₁₂-His-X₃₋₅-His (where X represents any amino acid) (Pabo and Sauer 1992). Crystal structure of a Zif268-DNA complex shows that the three zinc fingers are

arranged in a C-shaped structure that fits into the major groove of DNA. Each zinc finger motif consists of an antiparallel β -sheet and an α -helix. The β -sheet and the α -helix were held together by the zinc ion.

4) Steroid Receptor

Steroid receptors are intracellular receptors and initiate signal transduction for steroid hormones, including vitamin D, retinoids, thyroid hormones and other important hormones. The structure of the DNA-binding domain of steroid receptors have 8 conserved cysteine residues. In spite of the existence of two sets of 4 cysteines, the DNA-binding domain of steroid receptors does not form a pair of zinc fingers (Pabo and Sauer 1992). The crystal structure of the glucocorticoid receptor showed that it forms a dimer (Luisi, Xu et al. 1991). The first helix of each subunit binds to the major groove of DNA. The crystal structure also showed that the second major helix makes contact with the DNA backbone.

5) Helix-Loop-Helix

Regulatory proteins containing helix-loop-helix domain have important roles in development process (Jones 2004). A typical helix-loop-helix motif consists of 2 long α -helices connected by a short loop. In the example of MyoD structure, it is a homo-dimer. The first helix makes contact with the major groove of DNA. Usually, the two subunits make symmetric contacts with DNA, but some other proteins such as SREBP-1a recognize DNA-binding site asymmetrically (Parraga, Bellolell et al. 1998).

2. Mapping transcription factor binding sites

When the DNA-binding domain makes contact with DNA sequence, the side chain of the DNA-binding domain forms hydrogen bond with bases. This is the most important recognition mechanism of the binding site. This recognition leads to sequence-specific binding of a transcription factor to DNA. However, the interaction of transcription factor and DNA is complex (Pabo and Sauer 1992). Although the base contact is the most significant, backbone contacts are also important in sequence recognition. There is also no one-to-one mapping of a side chain residue to a certain base. For example, arginine has been found to contact with T/G base. As a result of the complexity and base-specificity of DNA recognition, not all bases in the transcription factor binding site are specific. Thus, transcription factors do not bind to only one sequence but a set of closely related sequences, with different affinities.

3. Structure and function of stem cell transcription factors and their binding sites

Some transcription factors were found to play crucial role during early embryonic stem cell development. A milestone study for the transcription network in stem cell identified three master transcription factors, Oct4, Sox2 and Nanog (Boyer, Lee et al. 2005, Loh, Wu et al. 2006, Chen, Xu et al. 2008).

Oct4 is a POU (PIT/OCT/UNC) family protein and it is a homeodomain transcription factor. In ES cells, Oct4 maintains the pluripotency of stem cell and prevent it from differentiation. To maintain pluripotency, the dosage of Oct4 is critical. It has been reported that loss of Oct4 led to differentiation into trophectoderm while a 2-fold increase of Oct4 results in differentiation into mixed population of mesodermal and endodermal cells (Niwa, Toyooka et al. 2005).

Oct4 also forms a heterodimer with Sox2 to regulate diverse gene targets. Sox2 is an HMG domain transcription factor, which contains 3 α -helices in structure. Sox2 knockout mice embryo failed to develop after implantation. Sox2-deficient ES cells were found to differentiate mainly into trophoctodermal lineages (Maruyama, Ichisaka et al. 2005). Similar to the effect of up-regulating Oct4, a 2-fold overexpression of Sox2 led ES cells to differentiate (Kopp, Ormsbee et al. 2008).

Nanog is also a homeodomain transcription factor. It forms a complex with Oct4 to regulate genes. Nanog knockout embryos fail to form epiblasts and become disorganized tissue (Chambers, Colby et al. 2003). It has been also shown that ES cells can self-renew in the absence of Nanog (Chambers, Silva et al. 2007). This implies that the main function of Nanog is stabilizing the pluripotent state of ES cells.

Another critical finding of the role of transcription factors in ES cells is shown by the generation of iPSCs (Takahashi and Yamanaka 2006). In this study, introducing transcription factors, Oct4, Sox2, Klf4 and Myc, was able to reprogram differentiated cells to pluripotent cells.

Transcription regulation network is critical to maintain the undifferentiated state of stem cell. As described above, a few important transcription factors are found to be sufficient to control the pluripotency of stem cell. These transcription factors regulate a broad spectrum of genes. Thus, understanding what targets they regulate becomes important. In the scenario of maintaining pluripotency, the number of related transcription factors is much smaller than that of the target genes. Thus, the target genes, or the binding site of these transcription factors should have some pattern.

Some transcription factors such as Oct4 and Sox2, Nanog and Smad1 were found to have highly similar binding sites (Chen, Xu et al. 2008). This is consistent with the fact that they form heterodimer to regulate gene transcription. Some other core transcription factors in ES cell were not found to have similar pattern in their binding sites (Loh, Wu et al. 2006). Although binding sites might be different, transcription factors may still work closely to regulate certain genes. Their specific recognition sequences do not necessarily map to target genes one-to-one. Thus, which gene or where on the genome these transcription factors bind to become more important and more informative.

Chen *et al.* systematically analyzed the binding loci of 13 core transcription factors in ES cell (Chen, Xu et al. 2008). Interestingly, two groups of transcription factors were found to have high co-occupancy of their binding loci on the genome. One group consists of Nanog, Sox2, Oct4 and Smad1. The other group consists of c-Myc, Zfx, n-Myc and E2f1. They suggested that the multiple transcription factor-binding loci from the Nanog centric cluster showed robust ES-cell-specific enhancer activity, compared to the Myc-centric cluster. And the transcriptional co-activator P300 showed high-co-occupancy of its binding sites with Nanog-centric cluster instead of Myc centric cluster (Faiola, Liu et al. 2005).

4. Localization of stem cell transcription factors by fluorescence microscopy

The transcription factors play important roles in regulating gene expression and thus governing cell differentiation. The interactions among core transcription factors in regulation of gene expression have been extensively studied with approaches combining

of traditional biochemistry, genomics and genetics (Levine and Tjian 2003, Chen, Xu et al. 2008, Consortium 2012, Levine, Cattoglio et al. 2014).

However, those approaches unable to provide spatial organization and temporal dynamics of the binding events of sequence-specific transcription factors, while fluorescence microscopy could partially overcome by providing a direct visualization and even bring the possibility of tracking the dynamics of transcription factors. The chromatin binding sites and localizations of the transcription factors have been investigated with different fluorescence microscopy techniques, such as fluorescence correlation spectroscopy (FCS), fluorescence recovery after photo-bleaching (FRAP) (Mazza, Abernathy et al. 2012), high-resolution imaging with aberration-corrected multi-focus microscopy (Abrahamsson, Chen et al. 2013), reflected light-sheet microscopy (RLSM)(Gebhardt, Suter et al. 2013, Liu, Legant et al. 2014), and total internal reflection fluorescence microscopy (TIRF)(Chen, Zhang et al. 2014). Chen, et al established their approach of single-cell/single-molecule imaging and tracking for the transcription factor binding dynamics. They announced that the assembly of Oct4 and Sox2 follows a hierarchically style, as Sox2 searching and binding to the specific DNA target prior to binding with Oct4 to form a stable complex (Chen, Zhang et al. 2014).

These analysis dedicate great efforts in developing assay procedure based on single-molecule/single-cell imaging and focusing on a certain or a few pair of transcription factors to gain a deep understanding and thorough investigation of the precise spatiotemporal regulation of gene expression by those core transcription factors.

Chen (Chen, et al. 2008) and Fang Fang (Fang, et al. 2014) have examined the binding profiles of 13 transcriptional factors in multiple transcription factor-binding loci using Chip-Seq technique. The binding profiles were shown in a heat map (Fig. 22). Colors in the heat map reflect the co-localization frequency of each pair of TFs in MTL. TFs have been clustered along both axes based on the similarity in their co-localization with other factors. Two groups of factors with similar binding profiles were observed: Nanog-centric cluster including Nanog, Sox2, Oct4 and Smad1; Myc-centric cluster including c-myc, n-myc, Zfx and E2F1. And furthermore, two transcription co-activators P300 and CBP were claimed to have similar binding profile with Nanog-centric cluster.

The similarities of the binding profiles of these two clusters of factors are investigated with biochemical techniques, however, by only looking at the data, we will not be able to answer the question: are these transcription factors bind to the same site at the same time? And the level of transcriptional factors is cell cycle dependent, while in the above studies the cells under analysis were in mixed populations from different cell cycle phases.

Therefore in this part, we aim to use high-resolution imaging tools to investigate and validate the co-localization between a set of key transcriptional factors.

Figure 22 Heat map of the binding profiles of 13 factors. Factors were clustered on both axes based on their binding profiles. Colors in the heat map reflect the co-localization frequency of each pair of TFs in MTL (yellow means more frequently co-localized, red means less) (from Chen, X et al. 2008).

4.3.2 Sample Preparation

1. Cell culture

E14 mouse ES cells, cultured under feeder-free conditions on surfaces coated with 0.1% gelatin, were maintained in Dulbecco's modified Eagle's medium (DMEM; GIBCO), supplemented with 15% heat-inactivated fetal bovine serum (FBS; GIBCO), 0.055 mM β -mercaptoethanol (GIBCO), 2 mM L-glutamine (GIBCO), 0.1 mM minimal essential medium with nonessential amino acids (GIBCO), and 1,000 U/ml of leukemia inhibitory factor (LIF) (Chemicon). Cells were pre-seeded on Ibidi 8 well u-slides with ibiTreat coverslip-like bottom for high-resolution imaging.

2. Cell synchronization

Cell synchronization methods include physical methods and chemical methods. Physical methods, such as mitotic shake off and flow cytometer, yielded small numbers of cell in each phase of cell cycle. So I choose to use chemical methods where chemical inhibitors are employed. The entire cell population is stalled at the same phase of the cell cycle.

Nocodazole was used for cell cycle synchronization. Nocodazole inhibits microtubule polymerization and thus prevents the formation of metaphase spindles. I have tested a range of concentrations and chose 50 ng/ml as the optimal concentration to ensure blocking and cell viability. However the blocked cells are less synchronized after treatment.

Thus I tested double blocking with thymidine and nocodazole, which could improve the synchronization rate. Thymidine treatment will block cells in G1/S phase, as excess

thymidine will induce feedback inhibition of DNA replication. The concentrations of thymidine are tested (0.5 mM-3 mM), and 1.25mM for 21hrs is chosen to ensure blocking rate and cell viability. After that, the cells are washed with 1X PBS 3 times, 5 mins each. By adding fresh medium, the revival cells start to release from the blocking and undergo the following cell cycle phases. The flow results are shown in Fig.23 and Fig. 24 for blocking and releasing.

Figure 23 flow profile of cell synchronization under different conditions. (A) Flow profile for control cell population. Two peaks were observed to show cell population in S and G2/M phases. (B) After treated with Thymidine, the cells are synchronized in S phase and the peak at G2/M phase is disappeared. (C) Treated with Nocodazole only. Cells were synchronized in G2/M phase. Peak for S phase disappeared. However, the cell survival rate decreased and the peak in G2/M phase was broad and will less cell number. (D) Double blocking with Thymidine followed by nocodazole. The cells are well synchronized in G2/M phase with a very sharp peak in the flow profile.

Figure 24 flow profile of cell releasing time points after synchronization. After double blocking, the drugs are washed out with PBS and cells are released from the blocking and went into the next cell cycle phase. After 3 hrs majority of cells went through the cell cycle and back to S phase.

Overall, I chose four time points (Fig. 25) to perform our analysis for the co-localization dynamics between factor pairs:

- 1) **S phase:** Blocking before DNA synthesis with thymidine.
- 2) **G2/M phase:** Blocking before spindle formation with Thymidine and Nocodazole double blocking.
- 3) **M phase:** Releasing 0.5 hour after Thymidine and Nocodazole blocking to let cells enter mitosis.
- 4) **G1 phase:** Releasing 3 hours after Thymidine and Nocodazole blocking to ensure most of the cells exit from mitosis.

Figure 25 Four time points chosen in the following analysis

3. Immuno-staining

Cells were fixed with 4% PFA for 20 minutes at exact time points described in 2.2, and followed by permeabilization with 0.3% Triton X100 for 10mins, washed with 1x PBS. Before staining with primary antibodies, cells are incubated with 5% BSA to block unspecific signals.

Nine transcription factors and co-activators (Oct4, Sox2, Nanog, P300, CBP, n-myc, c-myc, ZFX and E2F1) are studied. They can be clustered into two groups sharing similar genomic binding sites. The Nanog-centric cluster includes Oct4, Sox2, and Nanog. While co-activators P300 and CBP are reported as sharing similar binding sites with Nanog-centric cluster. The Myc-centric cluster includes n-myc, c-myc, ZFX and E2F1 (Chen, Xu et al. 2008, Fang, Xu et al. 2014). We would expect higher co-localization shown with the clusters and lower between clusters.

Primary antibodies for these 9 transcription factors and co-activators have been tested ahead to ensure the optimized concentrations (Appendix II). Primary antibody staining last for 1 hour at room temperature, and the samples are washed with 5% BSA/1X PBS three times. Alexa Fluor 488 and 568 were used as secondary antibodies, together with Hoechst33342 to label the nuclei or DNA.

4.3.3 Microscopic system and setting

The immune stained samples were imaged on a PerkinElmer UltraView Vox Spinning Disk system. An Olympus UPlan SApo 100x/1.4 oil objective was used for image collection. 3D stack of single ES cell colonies was taken. Sample thickness was 20 μm (10 μm away from focus plane), and z-resolution was 0.1 μm . (Fig. 26), and the double staining cell samples for each factor pair were shown in fig.27-29.

Figure 26 3D projections of image samples of 9 TFs. A single representative colony shown for each factor.

Figure 27 Double staining for each transcription factor pairs – Nanog-centric cluster

Figure 28 Double staining for each transcription factor pairs – between Nanog-centric cluster and Myc-centric cluster

Figure 29 Double staining for each transcription factor pairs –Myc-centric cluster

4.3.4 Image Processing

For each image collected, a bright field image would be taken with an auto-fluorescent slide for each channel. The bright images were used to correct the background uneven illumination for each image before further analysis.

Huygens Pro software was used to perform de-convolution to all the images with a theoretical PSF designed for UltraView Spinning Disk system.

Three-dimensional level set algorithm was employed for the cell segmentation, and the Nuclei channel was used as cell marker. The region occupied by TFs is segmented with automatic thresholding algorithm based on correlation coefficient, which has been described in previous chapter.

The cell region is divided as cytoplasm, nucleus and condensed DNA based on the intensities of Nuclei channel. The cells will then be categorized into normal cells and cells undergoing mitosis based on the shape and intensity distribution of the nucleus.

4.3.5 Results and Discussion

1. Transcriptional Factors in Embryonic Stem Cells

First, I measured the co-localization of all 9 factors (Fig. 30-33). In the heat map, the horizontal axis shown the percentage of one transcriptional factor co-localized with the other nine. As the co-occurrence of factor X with factor Y is not necessary the same as the one of factor Y with factor X, the values may not be symmetrical along the diagonal line. Due to the limit of host species of the antibodies, the co-localization data for some of the TF pairs are currently not available in this study.

The co-localizations within the Nanog-centric clusters (blue window) are significantly higher than the ones between two clusters, which is not significant within Myc-centric cluster (red window). It suggests that the interaction between TF pairs of Myc-centric cluster is not as active as Nanog-centric cluster. The highest co-occurrence is Oct4/Sox2, and the second highest is Oct4/CBP. And surprisingly, there is little co-occurrence between P300 and CBP, although they are reported as two highly homologous co-activators and play redundant roles in maintaining the undifferentiated state of mouse ES cells (Goodman and Smolik 2000, Chan and La Thangue 2001).

Figure 30 Co-localization heat map of TF pairs in S phase

Figure 31 Co-localization heat map of TF pairs in G2/M phase

Figure 32 Co-localization heat map of TF pairs in M phase

Figure 33 Co-localization heat map of TF pairs in G1 phase

2. Co-Localization of Transcriptional Factor Pairs

I then examine the dynamic co-localization of each factor with others during 4 cell-cycle time points. And the percentage of co-localization within Nucleus, condensed DNA, and cytoplasm are calculated separately. In order to compare the changes between time point 1 and time point 4, the bars of time point 1 are duplicated at the time point 5 positions (Fig. 34-42).

1) Oct4

The co-localization of Oct4 is shown in Fig. 34.

The overall co-localization of Oct4 is higher within Nanog-centric cluster, compared to Myc-centric cluster. And among factors of Nanog-centric cluster, Oct4 is highly co-localized with Sox2 and CBP. The co-localization of Oct4 with other factors is usually happened within nuclei region but not condensed DNA. And during mitosis phase, the co-localization of Oct4 with other factors is reduced, as they are not binding to the condensed DNA during that period.

Though literature reports, Oct4 and Sox2 together regulate miR-302 cluster, which will promote G1/S transition in hESCs (Card, Hebbar et al. 2008). Oct4, Sox2 and Nanog frequently bind to the sites that are closely localized in the chromatin (Boyer, Lee et al. 2005, Loh, Wu et al. 2006, Chambers and Tomlinson 2009). Beta-catenin/CBP/Oct4 complex could play an important role in maintaining the pluripotency in ES cells (Chew, Loh et al. 2005, Wagner, Xu et al. 2010, Kelly, Ng et al. 2011, Miki, Yasuda et al. 2011). Those literatures suggest Oct4 should co-localize with Sox2, Nanog, and CBP, which is partially proved by this analysis.

2) Sox2

The co-localization of Sox2 is shown in Fig. 35.

The overall co-localization of Sox2 is higher with Nanog-centric cluster, compared to Myc-centric cluster. And among factors of Nanog-centric cluster, Sox2 is highly co-localized with Oct4 and CBP. The co-localization of Sox2 with other factors is usually happened within nuclei region but not condensed DNA. One interesting founding is that Sox2 and CBP have a significant increasing co-localization during M phase, while others all decreased, suggesting that they might have an active interaction within the cytoplasm during M phase. Literature suggested that Sox2 should highly co-localize with Oct4, Nanog and P300/CBP (Card, Hebbar et al. 2008, Chen, Xu et al. 2008, Chambers and Tomlinson 2009), which is partially proved by this analysis.

3) Nanog

The co-localization of Nanog is shown in Fig. 36.

The overall co-localization of Nanog is higher with Nanog-centric cluster, compared to Myc-centric cluster. And among factors of Nanog-centric cluster, Nanog is well co-localized with Oct4 and P300/CBP. The co-localization of Nanog with other factors is usually happened within nuclei region but not condensed DNA. And during mitosis phase, the co-localization of Nanog with other factors is reduced, as they are not binding to the condensed DNA during that period. Literature suggested that Nanog should co-localize with Oct4, Sox2, and P300 (Boyer, Lee et al. 2005, Loh, Wu et al. 2006, Chambers and Tomlinson 2009), which is partially proved by this analysis.

Figure 34 Co-localization of Oct4

Figure 35 Co-localization of Sox2

Figure 36 Co-localization of Nanog

4) P300

The co-localization of P300 is shown in Fig. 37.

The overall co-localization of P300 is higher with the Nanog-centric cluster, compared to Myc-centric cluster. And among factors of Nanog-centric cluster, P300 is well co-localized with Oct4 and Sox2, which was suggested by literature (Chen, Xu et al. 2008, Baltus, Kowalski et al. 2009, Chambers and Tomlinson 2009). However, again similar as shown in Fig. 21, there is not much co-localization between P300 and CBP, although they are reported as two highly homologous co-activators and play redundant roles in maintaining the undifferentiated state of mouse ES cells (Goodman and Smolik 2000, Chan and La Thangue 2001). One possible explanation is that P300 and CBP play a redundant role in the network and they bind to the same site but not at the same time. During mitosis phase, the co-localizations of P300 with Nanog and CBP are increased with certain significance.

5) CBP

The co-localization of CBP is shown in Fig. 38.

The overall co-localization of CBP is higher with Nanog-centric cluster, compared to Myc-centric cluster. And it is highly co-localized with all other four factors of Nanog-centric cluster, a little bit less with P300. The co-localization of CBP with other factors is usually happened within nuclei region but not condensed DNA. Besides its interaction with Nanog-centric cluster, P300/CBP acetylase will stimulated the activity of E2F1 (Martínez - Balbás, Bauer et al. 2000). A relatively high co-localization of CBP with E2F1

with nuclear region is seen compared to the others in Myc-centric clusters. And again during mitosis phase, CBP is shown a very significant increasing in co-localization with Sox2 factor. The co-localization happened in the cytoplasm region but not on condensed DNA. Repeat experiments and examination with other methods are necessary for confirming and investigating this interesting phenomenon.

Figure 37 Co-localization of P300

Figure 38 Co-localization of CBP

6) N-myc

The co-localization of N-myc is shown in Fig. 39.

The overall co-localization of N-myc is relatively low with all nine factors, no matter Nanog-centric or Myc-centric. And the co-localizations of N-myc with other factors are usually happened in cytoplasm. The interactions of N-myc with other 8 factors is not well studied.

7) C-myc

The co-localization of C-myc is shown in Fig. 40.

C-myc shows relatively high co-localization with CBP within cytoplasm and N-myc within Nuclei, and ZFX within both cytoplasm and Nuclei. The co-localization of C-myc and ZFX goes up to more than 40% during mitosis phase. It is reported that C-myc together with ZFX and other two cofactors can form a network that regulates the gene transcription involved in cell cycle, cell death and cancer in mouse embryonic stem cells (Hu, Kim et al. 2009), which is partially proved by this analysis. The co-localization between C-myc and E2F1 is extremely low comparing to others. It is reported that C-myc and E2F1 can regulate cell growth and cell death. Overexpression may lead to apoptosis to the cells(Matsumura, Tanaka et al. 2003).

8) ZFX

The co-localization of ZFX is shown in Fig. 41.

Similar with the analysis of C-myc, compared to other factors, ZFX and C-myc have a high co-localization especially in M phase (up to 40%), suggesting that there could be significant high interaction between those two factors, especially during mitosis phase.

9) E2F1

The co-localization of E2F1 is shown in Fig. 42.

The co-localization of E2F1 appears low in average and mostly happened in cytoplasm. Compared to others, CBP has a higher interaction with E2F1, as being reported as E2F1 activity might be stimulated by p300/CBP acetylase (Martínez - Balbás, Bauer et al. 2000).

Figure 39 Co-localization of N-myc

Figure 40 Co-localization of C-myc

Figure 41 Co-localization of ZFX

Figure 42 Co-localization of E2F1

4.4 Conclusion

In this chapter, I tested the performances of several co-localization measurements with two sets of synthetic images generated with different methods. The test focuses on two aspects of co-localization: co-occurrence and correlation. By examining the performances of the co-localization measurements, we would gain clearer idea of their different capability in reporting the variations of co-occurrence and intensity correlation of different datasets.

And thus I propose a method to separate the aspects of co-occurrence and correlation from the co-localization quantification, which would be greatly helpful in providing biological meaningful explanations to the quantification results.

Furthermore, I apply this method in quantifying co-localization of transcription factors in mouse Embryonic Stem cells from high-resolution 3-dimensional images. The cells are treated before the analysis to represent different stages of cell cycle. And thus I am able to examine the dynamics of the interaction between TF pairs during cell cycle. With this analysis, I am trying to discover the possibility of using high-resolution imaging tools with image processing algorithm to investigate the relationship and interactions between TFs. With the characteristic of high-content and high-resolution imaging, the analysis would be quite useful in examine and filter out possible interactions between TFs, before time-consuming biological experiments can be applied. The interactions of TFs usually occur in genomic dimension level. Thus in the future, super-resolution is necessary to be incorporated into the analysis for more accurate results.

Chapter 5

Conclusion and Future Directions

5.1 Conclusion

The importance of subcellular localization of proteins and the interaction of two protein molecules occupying the same location is generally accepted as the subcellular localization and the bio-physiological functions of a certain protein are closely related. In this thesis work, we have shown the concise and efficient characteristic of fluorescence microscopic term co-localization in addressing this problem.

In this thesis work, I first introduce co-localization into the process of quantification and prediction of subcellular localization, and the corresponding computational frameworks are developed to perform and evaluate the performance of the co-localization in quantifying and predicting of subcellular localization. A new co-localization measurement is defined to estimate the degree of overlapping between protein and different subcellular compartments, which can be used to generate the co-localization profiles of proteins across different subcellular compartments. By using this single co-localization measurement, the co-localization profiles of proteins can be easily compared and clustered. Furthermore, the method provides the possibility in quantifying the dynamic changes of protein co-localization.

The co-localization measurement can either work as a single parameter to estimate the relationship between a specific protein and a series of subcellular compartments to generate a co-localization profile or work with many other statistic measurements serving as a set of co-localization features, which can be used to predict the subcellular localization itself or with other protein sequence features to improve the prediction accuracy. Although the size and quality of the dataset limits the prediction accuracy with the co-localization features, the analysis in this thesis has shown the capability of the co-localization features in prediction of subcellular localization and also provides a good complimentary to protein sequence features to increasing the prediction accuracy.

On the other hand, tested with synthetic images, I developed a method based on the co-localization measurement to estimate the co-occurrence and correlation separately, which would be greatly helpful in providing biological meaningful explanations to the quantification results, especially for mid-value results. The method was applied on a high-resolution and high-content 3D image set of a certain core transcription factors blocked at several time points of cell cycle. The dynamic analysis provides an insight in investigating the relationship between transcription factor pairs during cell cycle as those core factors may bind to the same or adjacent genome binding sites or form aggregates to regulate gene expression together. The high-throughput statistical analysis ensures the efficiency in filtering out the most interesting phenomenon from dozens of data points for further focused investigation.

5.2 Future Work and Directions

1. In order to establish a more comprehensive model of quantified subcellular

localizations of proteins, the computational framework needs to be applied on a larger protein library with higher resolution, thus to eliminate the bias generated by over-fit classification. Furthermore, as in this analysis I focused on cytoskeletal proteins, the actin network is used as a reference channel. For other protein families, more reference channels can be induced according to different analyzing interesting.

2. The limitation in the current image set also affect the prediction accuracy of the classifiers. Training with a more general protein library would largely help in investigating the feature performance in different classifiers, thus to infer the biological meaning hidden beneath the computational statistics.
3. The co-localization measurements separated in two aspects of co-occurrence and correlation have been tested and optimized on synthetic images, but not on real images. The more complex imaging environment would overlay different kinds of noises onto the original signal. Incorporation of some signal processing techniques, wavelet for example, into the framework would be useful in increasing the stability of the quantification and reducing the false positive co-localizations.
4. By examining carefully through the co-localization data of transcription factors, a few interesting phenomena has risen interesting in further investigation. We will focus on those certain transcription factor pairs with biological assays. The further investigation of co-localization between transcription factor and specific DNA binding sites or enhancers would also be interesting. The computational results

and experimental results could be used to confirm for each other, thus to draw noticeable conclusion. Super-resolution microscopy would also be powerful to incorporate into the analysis.

Appendix

I. Invitrogen GFP-tagged cytoskeletal protein constructs

Group	brief	name	function
First AB	DCAMKL1	doublecortin and CaM kinase-like 1	microtubule-associated protein. has microtubule polymerizing activity that is independent of its protein kinase activity .Probable kinase that may be involved in a calcium-signaling pathway controlling neuronal migration in the developing brain. May also participate in functions of the mature nervous system
	CORO2B	coronin, actin binding protein, 2B	May play a role in the reorganization of neuronal actin structure
	PDLIM3	PDZ and LIM domain 3	May play a role in the organization of actin filament arrays within muscle cells
	WASPIP	Wiskott-Aldrich syndrome protein interacting protein	May have direct activity on the actin cytoskeleton. Induces actin polymerization and redistribution. Contributes with NCK1 and GRB2 in the recruitment and activation of WASL. May participate in regulating the subcellular localization of WASL, resulting in the disassembly of stress fibers in favor of filopodia formation.
	KARCA1 (KLHDC9)	kelch/ankyrin repeat containing cyclin A1 interacting protein	
	CTTN	cortactin	The encoded protein is localized in the cytoplasm and in areas of the cell-substratum contacts. (1) regulating the interactions between components of adherens-type junctions and (2) organizing the cytoskeleton and cell adhesion structures of epithelia and carcinoma cells.
	VIL1	villin 1	a member of a family of calcium-regulated actin-binding proteins. This protein

			represents a dominant part of the brush border cytoskeleton which functions in the capping, severing, and bundling of actin filaments.
	ACTB	actin, beta	This gene encodes one of six different actin proteins. Actins are highly conserved proteins that are involved in cell motility, structure, and integrity. This actin is a major constituent of the contractile apparatus and one of the two nonmuscle cytoskeletal actins.
	WASL	Wiskott-Aldrich syndrome-like	The Wiskott-Aldrich syndrome (WAS) family of proteins share similar domain structure, and are involved in transduction of signals from receptors on the cell surface to the actin cytoskeleton. associate with the small GTPase, Cdc42, known to regulate formation of actin filaments, and the cytoskeletal organizing complex, Arp2/3.
	DNM2	dynamamin 2	Dynamins represent one of the subfamilies of GTP-binding proteins. associated with microtubules. They have been implicated in cell processes such as endocytosis and cell motility, and in alterations of the membrane that accompany certain activities. Dynamins bind many proteins that bind actin and other cytoskeletal proteins. Dynamins can also self-assemble, a process that stimulates GTPase activity. Microtubule-associated force-producing protein involved in producing microtubule bundles and able to bind and hydrolyze GTP. Most probably involved in vesicular trafficking processes, in particular endocytosis
	ZYX	zyxin	Focal adhesions are actin-rich structures that enable cells to adhere to the extracellular matrix and at which protein complexes involved in signal transduction assemble. Zyxin is a zinc-binding phosphoprotein that concentrates at focal adhesions and along the actin cytoskeleton. Zyxin may function as a messenger in the signal transduction pathway that mediates adhesion-stimulated changes in gene expression and may modulate the cytoskeletal organization of actin bundles.

	TUBA6	tubulin alpha 6	Tubulin is the major constituent of microtubules. It binds two moles of GTP, one at an exchangeable site on the beta chain and one at a non-exchangeable site on the alpha-chain
First CD	CDC42	cell division cycle 42	regulates signaling pathways that control diverse cellular functions including cell morphology, migration, endocytosis and cell cycle progression. This protein could regulate actin polymerization through its direct binding to Neural Wiskott-Aldrich syndrome protein (N-WASP) Plasma membrane-associated small GTPase which cycles between an active GTP-bound and an inactive GDP-bound state.
	CFL1	cofilin 1	widely distributed intracellular actin-modulating protein that binds and depolymerizes filamentous F-actin and inhibits the polymerization of monomeric G-actin in a pH-dependent manner. It is involved in the translocation of actin-cofilin complex from cytoplasm to nucleus. Controls reversibly actin polymerization and depolymerization in a pH-sensitive manner. It has the ability to bind G- and F-actin in a 1:1 ratio of cofilin to actin. It is the major component of intranuclear and cytoplasmic actin rods
	TGOLN2	trans-golgi network protein 2	May be involved in regulating membrane traffic to and from trans-Golgi network
	TAGLN	transgelin	The protein encoded by this gene is a transformation and shape-change sensitive actin cross-linking/gelling protein found in fibroblasts and smooth muscle. Involved in calcium interactions and contractile properties of the cell that may contribute to replicative senescence
	WAS	Wiskott-Aldrich syndrome	involved in transduction of signals from receptors on the cell surface to the actin cytoskeleton. associate with the small GTPase, Cdc42, known to regulate formation of actin filaments, and the cytoskeletal organizing complex, Arp2/3. a cytoplasmic protein. Effector protein for Rho-type GTPases, providing a link with the Arp2/3 complex that regulates the structure

			and dynamics of the actin cytoskeleton. Important for efficient actin polymerization.
	EVL	Enah/Vasp-like	actin-associated proteins involved in a range of processes dependent on cytoskeleton remodeling and cell polarity. enhances actin nucleation and polymerization.
	TUBE1	tubulin, epsilon 1	This beta-like tubulin localizes to the pericentriolar material in a cell cycle-dependent way. It first localizes to the older centrosome in a newly duplicated pair, but then associates with both.
	TNS	tensin	The protein encoded by this gene localizes to focal adhesions, regions of the plasma membrane where the cell attaches to the extracellular matrix. This protein crosslinks actin filaments and contains a Src homology 2 (SH2) domain, which is often found in molecules involved in signal transduction. May be involved in cell migration, cartilage development and in linking signal transduction pathways to the cytoskeleton
	CLIPR-59 (CLIP3)	CLIP-170-related 59 kDa protein	Functions as a cytoplasmic linker protein. Involved in TGN-endosome dynamics
	DSTN	destrin (actin depolymerizing factor)	belongs to the actin-binding proteins ADF family. This family of proteins is responsible for enhancing the turnover rate of actin in vivo. This gene encodes the actin depolymerizing protein that severs actin filaments (F-actin) and binds to actin monomers (G-actin)
	ITGB2	integrin beta 2	belongs to the integrin beta chain family of proteins. Integrins are integral cell-surface proteins composed of an alpha chain and a beta chain. Integrins are known to participate in cell adhesion as well as cell-surface mediated signalling.
	CORO1B	coronin, actin binding protein, 1B	WD repeat-containing actin-binding proteins that regulate cell motility. Regulates leading edge dynamics and cell motility in fibroblasts. May be involved in cytokinesis and signal transduction

First EF	ATP6V1C1	ATPase, H ⁺ transporting, lysosomal 42kDa, V1 subunit C	Subunit C is necessary for the assembly of the catalytic sector of the enzyme and is likely to have a specific function in its catalytic activity. V-ATPase is responsible for acidifying a variety of intracellular compartments in eukaryotic cells
	TAGLN3	transgelin 3	
	VIL2	villin 2 (ezrin)	The cytoplasmic peripheral membrane protein encoded by this gene functions as a protein-tyrosine kinase substrate in microvilli. As a member of the ERM protein family, this protein serves as an intermediate between the plasma membrane and the actin cytoskeleton. This protein plays a key role in cell surface structure adhesion, migration and organization, and it has been implicated in various human cancers.
	TUBD1	tubulin, delta 1	In the elongating spermatid it is associated with the manchette, a specialized microtubule system present during reshaping of the sperm head
	ATP5G2	ATP synthase, H ⁺ transporting, mitochondrial F0 complex, subunit C2 (subunit 9)	Mitochondrial membrane ATP synthase (F(1)F(O) ATP synthase or Complex V) produces ATP from ADP in the presence of a proton gradient across the membrane which is generated by electron transport complexes of the respiratory chain. F-type ATPases consist of two structural domains, F(1) - containing the extramembraneous catalytic core and F(0) - containing the membrane proton channel, linked together by a central stalk and a peripheral stalk.
	TUBGCP3	tubulin, gamma complex associated protein 3	Gamma-tubulin complex is necessary for microtubule nucleation at the centrosome
	PARVG	parvin, gamma	actin-binding proteins associated with focal contacts. Probably plays a role in the regulation of cell adhesion and cytoskeleton organization
	RDX	radixin	cytoskeletal protein that may be important in linking actin to the plasma membrane. It is highly similar in sequence to both ezrin and moesin.

	ITGB3BP	integrin beta 3 binding protein (beta3-endonexin)	Transcription coregulator that can have both coactivator and corepressor functions. Isoform 1 is involved in the coactivation of nuclear receptors for retinoid X (RXRs) and thyroid hormone (TRs) in a ligand-dependent fashion. In contrast, it does not coactivate nuclear receptors for retinoic acid, vitamin D, progesterone receptor, nor glucocorticoid. Acts as a coactivator for estrogen receptor alpha. Acts as a transcriptional corepressor via its interaction with the NFKB1 NF-kappa-B subunit, possibly by interfering with the transactivation domain of NFKB1.
	FSCN3	fascin 3	Acts as an actin bundling protein
	TEKT3	tektin 3	filament-forming proteins that are coassembled with tubulins to form ciliary and flagellar microtubules. Forms filamentous polymers in the walls of ciliary and flagellar microtubules
	PAK4	p21 protein (Cdc42/Rac)-activated kinase 4	Activates the JNK pathway. Plays a role in the reorganization of the actin cytoskeleton and in the formation of filopodia. Phosphorylates and inactivates the protein phosphatase SSH1, leading to increased inhibitory phosphorylation of the actin binding/depolymerizing factor cofilin. Decreased cofilin activity may lead to stabilization of actin filaments.
First GH	GJB2	gap junction protein beta 2	The gap junctions were first characterized by electron microscopy as regionally specialized structures on plasma membranes of contacting adherent cells. These structures were shown to consist of cell-to-cell channels that facilitate the transfer of ions and small molecules between cells. The gap junction proteins, also known as connexins, purified from fractions of enriched gap junctions from different tissues differ.
	KALPHA1 (TUBA1B)	tubulin, alpha 1b	Tubulin is the major constituent of microtubules. It binds two moles of GTP, one at an exchangeable site on the beta chain and one at a non-exchangeable site on the alpha-chain

	CAPZA3	capping protein (actin filament) muscle Z-line, alpha 3	actin capping protein, one of the F-actin capping protein alpha subunit family. bind in a Ca(2+)-independent manner to the fast growing ends of actin filaments (barbed end) thereby blocking the exchange of subunits at these ends. Unlike other capping proteins (such as gelsolin and severin), these proteins do not sever actin filaments. May play a role in the morphogenesis of spermatid
	ACTG1	actin, gamma 1	Actins are highly conserved proteins that are involved in various types of cell motility, and maintenance of the cytoskeleton. Actin, gamma 1, encoded by this gene, is a cytoplasmic actin found in nonmuscle cells.
	KPTN	kaptin (actin binding protein)	May be involved in actin dynamics. May play a role in producing the sensory apparatus in hair cells. May play a role in actin rearrangements that accompany platelet activation and stereocilia formation
	TEKT1	tektin 1	This gene is predominantly expressed in the testis and in mouse, tektin 1 mRNA was localized to the spermatocytes and round spermatids in the seminiferous tubules, indicating that it may play a role in spermatogenesis.
	PTK9 (TWF1)	twinfilin, actin-binding protein, homolog 1 (Drosophila)	Actin-binding protein involved in motile and morphological processes. Inhibits actin polymerization, likely by sequestering G-actin. By capping the barbed ends of filaments, it also regulates motility. Seems to play an important role in clathrin-mediated endocytosis and distribution of endocytic organelles
	KIF3A	kinesin family member 3A	Microtubule-based anterograde translocator for membranous organelles. Plus end-directed microtubule sliding activity in vitro
	TPM1	tropomyosin 1 (alpha)	Binds to actin filaments in muscle and non-muscle cells. Plays a central role, in association with the troponin complex, in the calcium dependent regulation of vertebrate striated muscle contraction. Smooth muscle contraction is regulated by interaction with caldesmon. In non-muscle cells is implicated in stabilizing cytoskeleton actin filaments

	TUBB2	tubulin, beta 2A	Tubulin is the major constituent of microtubules. It binds two moles of GTP, one at an exchangeable site on the beta chain and one at a non-exchangeable site on the alpha-chain
	ITGB1	integrin beta 1	Integrin family members are membrane receptors involved in cell adhesion and recognition in a variety of processes including embryogenesis, hemostasis, tissue repair, immune response and metastatic diffusion of tumor cells. This gene encodes a beta subunit.
	ADAM15	ADAM metallopeptidase domain 15	ADAM family members are type I transmembrane glycoproteins known to be involved in cell adhesion and proteolytic ectodomain processing of cytokines and adhesion molecules. this protein may function in cell-cell adhesion as well as in cellular signaling. May be involved in cell-surface proteolysis, cell adhesion or intracellular protein maturation
Last AB	PLS1	plastin 1	Actin-bundling protein in the absence of calcium
	VASP	vasodilator-stimulated phosphoprotein	Ena/VASP proteins are actin-associated proteins involved in a range of processes dependent on cytoskeleton remodeling and cell polarity in migrating cells. VASP promotes actin nucleation and increases the rate of actin polymerization in the presence of capping protein. Plays a role in actin-based activity of Listeria monocytogenes in platelets
	HTATIP (KAT5)	K(lysine) acetyltransferase 5	This protein is a histone acetylase that has a role in DNA repair and apoptosis and is thought to play an important role in signal transduction.
	ARHGEF6	Rac/Cdc42 guanine nucleotide exchange factor (GEF) 6	The encoded protein belongs to a family of cytoplasmic proteins that activate the Ras-like family of Rho proteins by exchanging bound GDP for GTP. Acts as a RAC1 guanine nucleotide exchange factor (GEF)
	WASF3	WAS protein family, member 3	The gene product is a protein that forms a multiprotein complex that links receptor kinases and actin. Binding to actin occurs through a C-terminal verprolin homology

			domain in all family members. The multiprotein complex serves to transduce signals that involve changes in cell shape, motility or function.
	KIAA0555 (JAMIP2)	janus kinase and microtubule interacting protein 2	
	EIF2C1	eukaryotic translation initiation factor 2C, 1	This gene encodes a member of the Argonaute family of proteins which play a role in RNA interference. This gene is located on chromosome 1 in a cluster of closely related family members including argonaute 3, and argonaute 4. Plays an important role in the eukaryotic peptide chain initiation process
Last CD	LPXN	leupaxin	It may function in cell type-specific signaling by associating with PYK2. As a substrate for a tyrosine kinase in lymphoid cells, this protein may also function in, and be regulated by tyrosine kinase activity.
	ATP1B3	ATPase, Na ⁺ /K ⁺ transporting, beta 3 polypeptide	This is the non-catalytic component of the active enzyme, which catalyzes the hydrolysis of ATP coupled with the exchange of Na(+) and K(+) ions across the plasma membrane. an integral membrane protein responsible for establishing and maintaining the electrochemical gradients of Na and K ions across the plasma membrane.
	VAMP4	vesicle-associated membrane protein 4	a member of the vesicle-associated membrane protein (VAMP)/synaptobrevin family. This protein may play a role in trans-Golgi network-to-endosome transport.
	PXN	paxillin	Cytoskeletal protein involved in actin-membrane attachment at sites of cell adhesion to the extracellular matrix
	DCTN1	Dynactin 1	Dynactin binds to both microtubules and cytoplasmic dynein. Dynactin is involved in a diverse array of cellular functions, including ER-to-Golgi transport, the centripetal movement of lysosomes and endosomes, spindle formation, chromosome movement, nuclear positioning, and axonogenesis. Required for the cytoplasmic dynein-driven retrograde movement of vesicles and organelles along microtubules. Dynein-dynactin interaction is a key

			component of the mechanism of axonal transport of vesicles and organelles
	WASF2	WAS protein family, member 2	The gene product is a protein that forms a multiprotein complex that links receptor kinases and actin. The multiprotein complex serves to transduce signals that involve changes in cell shape, motility or function. Downstream effector molecules involved in the transmission of signals from tyrosine kinase receptors and small GTPases to the actin cytoskeleton
	MSN	moesin--Membrane-organizing extension spike protein	function as cross-linkers between plasma membranes and actin-based cytoskeletons. Moesin is localized to filopodia and other membranous protrusions that are important for cell-cell recognition and signaling and for cell movement.
	KIF2C	kinesin family member 2C	a member of kinesin-like protein family. Proteins of this family are microtubule-dependent molecular motors that transport organelles within cells and move chromosomes during cell division. This protein is important for anaphase chromosome segregation and may be required to coordinate the onset of sister centromere separation. Present throughout the cell cycle, associates with centromeres at early prophase, and remains associated with the centromere until after telophase
	ITGB7	integrin, beta 7	Integrin beta-7 (Peyer patches-specific homing receptor LPAM-1) is an adhesion molecule that mediates lymphocyte migration and homing to gut-associated lymphoid tissue (GALT). Integrin beta-7 interacts with the cell surface adhesion molecules MADCAM1 which is normally expressed by the vascular endothelium of the gastrointestinal tract. Interacts also with VCAM1 and fibronectin, an extracellular matrix component.
	CAV3	caveolin 3	functions as a component of the caveolae plasma membranes found in most cell types. Caveolin proteins are proposed to be

			scaffolding proteins for organizing and concentrating certain caveolin-interacting molecules
Last EF	PLS3	plastin 3 (T isoform)	a family of actin-binding proteins that are conserved throughout eukaryote evolution and expressed in most tissues of higher eukaryotes. Actin-bundling protein found in intestinal microvilli, hair cell stereocilia, and fibroblast filopodia
	ADRM1	adhesion regulating molecule 1	an integral plasma membrane protein which promotes cell adhesion. The encoded protein is thought to undergo O-linked glycosylation. Functions as a proteasomal ubiquitin receptor
	ARPC1B	actin related protein 2/3 complex, subunit 1B, 41kDa	Functions as component of the Arp2/3 complex which is involved in regulation of actin polymerization and together with an activating nucleation-promoting factor (NPF) mediates the formation of branched actin networks
	TUBG1	tubulin, gamma 1	Tubulin is the major constituent of microtubules. Gamma tubulin is found at microtubule organizing centers (MTOC) such as the spindle poles or the centrosome, suggesting that it is involved in the minus-end nucleation of microtubule assembly
	MRLC2	myosin regulatory light chain MRLC2	Myosin regulatory subunit that plays an important role in regulation of both smooth muscle and nonmuscle cell contractile activity via its phosphorylation. Implicated in cytokinesis, receptor capping, and cell locomotion
	GTSE1	G-2 and S-phase expressed 1	is only expressed in the S and G2 phases of the cell cycle, where it colocalizes with cytoplasmic tubulin and microtubules. May be involved in p53-induced cell cycle arrest in G2/M phase by interfering with microtubule rearrangements that are required to enter mitosis. Overexpression delays G2/M phase progression
	TAGLN2	transgelin 2	one of the earliest markers of differentiated smooth muscle.
	ARPC5	actin related protein 2/3 complex subunit 5	Functions as component of the Arp2/3 complex which is involved in regulation of actin polymerization and together with an activating nucleation-promoting factor

			(NPF) mediates the formation of branched actin networks
	ACTN2	actinin, alpha 2	Alpha actinin is an actin-binding protein with multiple roles in different cell types. In contrast, skeletal, cardiac, and smooth muscle isoforms are localized to the Z-disc and analogous dense bodies, where they help anchor the myofibrillar actin filaments. This gene encodes a muscle-specific, alpha actinin isoform that is expressed in both skeletal and cardiac muscles. F-actin cross-linking protein which is thought to anchor actin to a variety of intracellular structures.
	MYO3A	myosin IIIA	Probable actin-based motor with a protein kinase activity. Probably plays a role in vision and hearing
	ACTN1	actinin, alpha 1	Alpha actinin is an actin-binding protein with multiple roles in different cell types. In nonmuscle cells, the cytoskeletal isoform is found along microfilament bundles and adherens-type junctions, where it is involved in binding actin to the membrane. This gene encodes a nonmuscle, cytoskeletal, alpha actinin isoform and maps to the same site as the structurally similar erythroid beta spectrin gene.
	FILIP1	filamin A interacting protein 1	By acting through a filamin-A/F-actin axis, it controls the start of neocortical cell migration from the ventricular zone. May be able to induce the degradation of filamin-A
Last GH	ARP3BETA (ARP11)	actin-related Arp11	May play a role in the suppression of metastatic potential in lung adenoma carcinoma cells
	CNN3	calponin 3, acidic	This encoded protein is associated with the cytoskeleton but is not involved in contraction. Thin filament-associated protein that is implicated in the regulation and modulation of smooth muscle contraction. It is capable of binding to actin, calmodulin, troponin C and tropomyosin. The interaction of calponin with actin inhibits the actomyosin Mg-ATPase activity

	LCP1	lymphocyte cytosolic protein 1 (L-plastin)	Actin-binding protein found in intestinal microvilli, hair cell stereocilia, and fibroblast filopodia. family of actin-binding proteins that are conserved throughout eukaryote evolution and expressed in most tissues of higher eukaryotes. The L isoform is expressed only in hemopoietic cell lineages. However, L-plastin has been found in many types of malignant human cells of non-hemopoietic origin suggesting that its expression is induced accompanying tumorigenesis in solid tissues.
	VIM	vimentin	Vimentins are class-III intermediate filaments found in various non-epithelial cells, especially mesenchymal cells
	NINJ2	ninjurin 2	Cell adhesion molecules mediate cell-to-cell and cell-to-extracellular matrix interactions during development, differentiation, and regeneration of the peripheral nervous system. NINJ2 is upregulated in Schwann cells in the distal nerve segment after peripheral nerve injury and can promote neurite outgrowth.
	PFN2	profilin 2	ubiquitous actin monomer-binding protein belonging to the profilin family. It is thought to regulate actin polymerization in response to extracellular signals. Binds to actin and affects the structure of the cytoskeleton. At high concentrations, profilin prevents the polymerization of actin, whereas it enhances it at low concentrations.
	PARVA	parvin, alpha	actin-binding proteins associated with focal contacts. Probably plays a role in the regulation of cell adhesion and cytoskeleton organization
	TPM2	tropomyosin 2 (beta)	Binds to actin filaments in muscle and non-muscle cells. Plays a central role, in association with the troponin complex, in the calcium dependent regulation of vertebrate striated muscle contraction. Smooth muscle contraction is regulated by interaction with caldesmon. In non-muscle cells is implicated in stabilizing cytoskeleton actin filaments
	MYO1A	myosin IA	Myosins are molecular motors that, upon interaction with actin filaments, utilize

			energy from ATP hydrolysis to generate mechanical force. Involved in directing the movement of organelles along actin filaments
	ANLN	anillin, actin binding protein	Required for cytokinesis. Essential for the structural integrity of the cleavage furrow and for completion of cleavage furrow ingression
	FSCN1	fascin homolog 1, actin-bundling protein	Organizes filamentous actin into bundles with a minimum of 4.1:1 actin/fascin ratio. Probably involved in the assembly of actin filament bundles present in microspikes, membrane ruffles, and stress fibers
	KRT8	keratin 8	This gene is a member of the type II keratin family. Type I and type II keratins heteropolymerize to form intermediate-sized filaments in the cytoplasm of epithelial cells. This protein plays a role in maintaining cellular structural integrity and also functions in signal transduction and cellular differentiation.

II. Primary antibodies for transcriptional factors

Antibody	Company	Cat. No.	Concentration
Oct4	Santa Cruz	Sc-8628	1:200-500
	Santa Cruz	Sc-9081	1:200
Sox2	Santa Cruz	Sc-17320	1:200-300
	Santa Cruz	Sc-20088	1:50
P300	Santa Cruz	Sc-585	1:50-200
CBP	Santa Cruz	Sc-583	1:50-100
	abcam	Ab50702	1:10
Nanog	R&D	AF-2729	1:100
c-myc	abcam	Ab32072	1:50-500
	abcam	Ab32	1:50-100
n-myc	Santa cruz	Sc-22836	1:50-100
ZFX	Santa Cruz	Sc-83959	1:50-500
E2F1	Santa Cruz	Sc-22820	1:50-200
	Santa Cruz	Sc-11762	1:50-100

References

Abeel, T., T. Helleputte, Y. Van de Peer, P. Dupont and Y. Saeys (2010). "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods." Bioinformatics **26**(3): 392-398.

Abrahamsson, S., J. Chen, B. Hajj, S. Stallinga, A. Y. Katsov, J. Wisniewski, G. Mizuguchi, P. Soule, F. Mueller and C. D. Darzacq (2013). "Fast multicolor 3D imaging using aberration-corrected multifocus microscopy." Nature methods **10**(1): 60-63.

Adler, J., S. Pagakis and I. Parmryd (2008). "Replicate- based noise corrected correlation for accurate measurements of colocalization." Journal of microscopy **230**(1): 121-133.

Adler, J. and I. Parmryd (2010). "Quantifying Colocalization by Correlation: The Pearson Correlation Coefficient is Superior to the Mander's Overlap Coefficient." Cytometry Part A **77A**(8): 733-742.

Affolter, M., A. Percival-Smith, M. Muller, W. Leupin and W. J. Gehring (1990). "DNA binding properties of the purified Antennapedia homeodomain." Proc Natl Acad Sci U S A **87**(11): 4093-4097.

Alberts, B. (2008). Molecular Biology of the Cell: Reference Edition, Garland Science.

Alberts, B., D. Bray, J. Lewis, M. Raff, K. Roberts and J. D. Watson (1994). "Molecular biology of the cell, 1994." Garland, New York: 139-194.

Baatz, M. and A. Schäpe (2000). "Multiresolution segmentation: an optimization approach for high quality multi-scale image segmentation." Angewandte Geographische Informationsverarbeitung XII: 12-23.

Baltus, G. A., M. P. Kowalski, H. Zhai, A. V. Tutter, D. Quinn, D. Wall and S. Kadam (2009). "Acetylation of sox2 induces its nuclear export in embryonic stem cells." Stem Cells **27**(9): 2175-2184.

Blobel, G. (1980). "Intracellular protein topogenesis." Proceedings of the National Academy of Sciences **77**(3): 1496-1500.

Boland, M. V., M. K. Markey and R. F. Murphy (1998). "Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images." Cytometry **33**(3): 366-375.

Boland, M. V. and R. F. Murphy (2001). "A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells." Bioinformatics **17**(12): 1213-1223.

Boutté, Y., M.-T. Crosnier, N. Carraro, J. Traas and B. Satiat-Jeunemaitre (2006). "The plasma membrane recycling pathway and cell polarity in plants: studies on PIN proteins." Journal of cell science **119**(7): 1255-1265.

Boyer, L. A., T. I. Lee, M. F. Cole, S. E. Johnstone, S. S. Levine, J. P. Zucker, M. G. Guenther, R. M. Kumar, H. L. Murray, R. G. Jenner, D. K. Gifford, D. A. Melton, R. Jaenisch and R. A. Young (2005). "Core transcriptional regulatory circuitry in human embryonic stem cells." Cell **122**(6): 947-956.

Bray, D. (2001). Cell Movements: From Molecules to Motility, Taylor & Francis Press.

Brennan, R. G. (1993). "The winged-helix DNA-binding motif: another helix-turn-helix takeoff." Cell **74**(5): 773-776.

Brennan, R. G. and B. W. Matthews (1989). "The helix-turn-helix DNA binding motif." J Biol Chem **264**(4): 1903-1906.

Brivanlou, A. H. and J. E. Darnell, Jr. (2002). "Signal transduction and the control of gene expression." Science **295**(5556): 813-818.

Card, D. A. G., P. B. Hebbar, L. Li, K. W. Trotter, Y. Komatsu, Y. Mishina and T. K. Archer (2008). "Oct4/Sox2-regulated miR-302 targets cyclin D1 in human embryonic stem cells." Molecular and cellular biology **28**(20): 6426-6438.

Chambers, I., D. Colby, M. Robertson, J. Nichols, S. Lee, S. Tweedie and A. Smith (2003). "Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells." Cell **113**(5): 643-655.

Chambers, I., J. Silva, D. Colby, J. Nichols, B. Nijmeijer, M. Robertson, J. Vrana, K. Jones, L. Grotewold and A. Smith (2007). "Nanog safeguards pluripotency and mediates germline development." Nature **450**(7173): 1230-1234.

Chambers, I. and S. R. Tomlinson (2009). "The transcriptional foundation of pluripotency." Development **136**(14): 2311-2322.

Chan, H. M. and N. B. La Thangue (2001). "p300/CBP proteins: HATs for transcriptional bridges and scaffolds." J Cell Sci **114**(Pt 13): 2363-2373.

Chen, J., Z. Zhang, L. Li, B.-C. Chen, A. Revyakin, B. Hajj, W. Legant, M. Dahan, T. Lionnet and E. Betzig (2014). "Single-molecule dynamics of enhanceosome assembly in embryonic stem cells." Cell **156**(6): 1274-1285.

Chen, S.-C., T. Zhao, G. J. Gordon and R. F. Murphy (2007). "Automated image analysis of protein localization in budding yeast." Bioinformatics **23**(13): i66-i71.

Chen, X., H. Xu, P. Yuan, F. Fang, M. Huss, V. B. Vega, E. Wong, Y. L. Orlov, W. Zhang, J. Jiang, Y. H. Loh, H. C. Yeo, Z. X. Yeo, V. Narang, K. R. Govindarajan, B. Leong, A. Shahab, Y. Ruan, G. Bourque, W. K. Sung, N. D. Clarke, C. L. Wei and H. H.

Ng (2008). "Integration of external signaling pathways with the core transcriptional network in embryonic stem cells." Cell **133**(6): 1106-1117.

Cheng, J. and J. C. Rajapakse (2009). "Segmentation of clustered nuclei with shape markers and marking function." Biomedical Engineering, IEEE Transactions on **56**(3): 741-748.

Chew, J.-L., Y.-H. Loh, W. Zhang, X. Chen, W.-L. Tam, L.-S. Yeap, P. Li, Y.-S. Ang, B. Lim and P. Robson (2005). "Reciprocal transcriptional regulation of Pou5f1 and Sox2 via the Oct4/Sox2 complex in embryonic stem cells." Molecular and cellular biology **25**(14): 6031-6046.

Claros, M. G., S. Brunak and G. v. Heijne (1997). "Prediction of N-terminal protein sorting signals." Current opinion in structural biology **7**(3): 394-398.

Coelho, L. P., T. Peng and R. F. Murphy (2010). "Quantifying the distribution of probes between subcellular locations using unsupervised pattern unmixing." Bioinformatics **26**(12): i7-i12.

Consortium, E. P. (2012). "An integrated encyclopedia of DNA elements in the human genome." Nature **489**(7414): 57-74.

Cortes, C. and V. Vapnik (1995). "Support-vector networks." Machine Learning **20**(3): 273-297.

Costes, S. V., D. Daelemans, E. H. Cho, Z. Dobbin, G. Pavlakis and S. Lockett (2004). "Automatic and quantitative measurement of protein-protein colocalization in live cells." Biophysical journal **86**(6): 3993-4003.

Damania, D., H. Subramanian, A. K. Tiwari, Y. Stypula, D. Kunte, P. Pradhan, H. K. Roy and V. Backman (2010). "Role of cytoskeleton in controlling the disorder strength of cellular nanoscale architecture." Biophysical journal **99**(3): 989-996.

Faiola, F., X. Liu, S. Lo, S. Pan, K. Zhang, E. Lyman, A. Farina and E. Martinez (2005). "Dual regulation of c-Myc by p300 via acetylation-dependent control of Myc protein turnover and coactivation of Myc-induced transcription." Molecular and cellular biology **25**(23): 10220-10234.

Fang, F., Y. Xu, K. K. Chew, X. Chen, H. h. Ng and P. Matsudaira (2014). "Coactivators p300 and CBP maintain the identity of mouse embryonic stem cells by mediating long-range chromatin structure." Stem Cells.

Farkash-Amar, S., E. Eden, A. Cohen, N. Geva-Zatorsky, L. Cohen, R. Milo, A. Sigal, T. Danon and U. Alon (2012). "Dynamic Proteomics of Human Protein Level and Localization across the Cell Cycle." PloS one **7**(11): e48722.

Finney, M. (1990). "The homeodomain of the transcription factor LF-B1 has a 21 amino acid loop between helix 2 and helix 3." Cell **60**(1): 5-6.

- Frixione, E. (2000). "Recurring views on the structure and function of the cytoskeleton: a 300-year epic." Cell motility and the cytoskeleton **46**(2): 73-94.
- Gebhardt, J. C. M., D. M. Suter, R. Roy, Z. W. Zhao, A. R. Chapman, S. Basu, T. Maniatis and X. S. Xie (2013). "Single-molecule imaging of transcription factor binding to DNA in live mammalian cells." Nature methods **10**(5): 421-426.
- Gehring, W. J., Y. Q. Qian, M. Billeter, K. Furukubo-Tokunaga, A. F. Schier, D. Resendez-Perez, M. Affolter, G. Otting and K. Wuthrich (1994). "Homeodomain-DNA recognition." Cell **78**(2): 211-223.
- Goodman, R. H. and S. Smolik (2000). "CBP/p300 in cell growth, transformation, and development." Genes Dev **14**(13): 1553-1577.
- Guyon, I., J. Weston, S. Barnhill and V. Vapnik (2002). "Gene selection for cancer classification using support vector machines." Machine learning **46**(1-3): 389-422.
- Han, G. S., V. Anh, A. P. Krishnajith and Y.-C. Tian (2013). "An Ensemble Method for Predicting Subnuclear Localizations from Primary Protein Structures." PloS one **8**(2): e57225.
- Hanley, J. A. (1982). "Characteristic (ROC) Curvel." Radiology **743**: 29-36.
- Harrison, S. C. and A. K. Aggarwal (1990). "DNA recognition by proteins with the helix-turn-helix motif." Annu Rev Biochem **59**: 933-969.
- Hu, G., J. Kim, Q. Xu, Y. Leng, S. H. Orkin and S. J. Elledge (2009). "A genome-wide RNAi screen identifies a new transcriptional module required for self-renewal." Genes & development **23**(7): 837-848.
- Hu, Y. and R. F. Murphy (2004). "Automated interpretation of subcellular patterns from immunofluorescence microscopy." Journal of immunological methods **290**(1): 93-105.
- Hua, S. and Z. Sun (2001). "Support vector machine approach for protein subcellular localization prediction." Bioinformatics **17**(8): 721-728.
- Huang, K. and R. F. Murphy (2004). "Boosting accuracy of automated classification of fluorescence microscope images for location proteomics." Bmc Bioinformatics **5**(1): 78.
- Hung, M.-C. and W. Link (2011). "Protein localization in disease and therapy." Journal of Cell Science **124**(20): 3381-3392.
- Janmey, P. A. (1998). "The cytoskeleton and cell signaling: component localization and mechanical coupling." Physiological reviews **78**(3): 763-781.
- Jaskolski, F., C. Mulle and O. J. Manzoni (2005). "An automated method to quantify and visualize colocalized fluorescent signals." Journal of neuroscience methods **146**(1): 42-49.

Jones, S. (2004). "An overview of the basic helix-loop-helix proteins." Genome Biol **5**(6): 226.

Kelly, K. F., D. Y. Ng, G. Jayakumaran, G. A. Wood, H. Koide and B. W. Doble (2011). " β -catenin enhances Oct-4 activity and reinforces pluripotency through a TCF-independent mechanism." Cell stem cell **8**(2): 214-227.

Khurana, S. (2006). Aspects of the Cytoskeleton: Advances in Molecular and Cell Biology. Advances in Molecular and Cell Biology. S. Khurana, Elsevier Science. **37**: 384 pages.

Kopp, J. L., B. D. Ormsbee, M. Desler and A. Rizzino (2008). "Small increases in the level of Sox2 trigger the differentiation of mouse embryonic stem cells." Stem Cells **26**(4): 903-911.

Lachmanovich, E., D. Shvartsman, Y. Malka, C. Botvin, Y. Henis and A. Weiss (2003). "Co-localization analysis of complex formation among membrane proteins by computerized fluorescence microscopy: application to immunofluorescence co-patching studies." Journal of microscopy **212**(2): 122-131.

Latchman, D. S. (1997). "Transcription factors: an overview." Int J Biochem Cell Biol **29**(12): 1305-1312.

Leonardy, S., G. Freymark, S. Hebener, E. Ellehauge and L. Sogaard-Andersen (2007). "Coupling of protein localization and cell movements by a dynamically localized response regulator in *Myxococcus xanthus*." The EMBO journal **26**(21): 4433-4444.

Levine, M., C. Cattoglio and R. Tjian (2014). "Looping back to leap forward: transcription enters a new era." Cell **157**(1): 13-25.

Levine, M. and R. Tjian (2003). "Transcription regulation and animal diversity." Nature **424**(6945): 147-151.

Li, Q., A. Lau, T. J. Morris, L. Guo, C. B. Fordyce and E. F. Stanley (2004). "A syntaxin 1, G α (o), and N-type calcium channel complex at a presynaptic nerve terminal: Analysis by quantitative immunocolocalization." Journal of Neuroscience **24**(16): 4070-4081.

Li, Q., A. Lau, T. J. Morris, L. Guo, C. B. Fordyce and E. F. Stanley (2004). "A syntaxin 1, Gao, and N-type calcium channel complex at a presynaptic nerve terminal: analysis by quantitative immunocolocalization." The Journal of neuroscience **24**(16): 4070-4081.

Liu, Z., W. R. Legant, B.-C. Chen, L. Li, J. B. Grimm, L. D. Lavis, E. Betzig and R. Tjian (2014). "3D imaging of Sox2 enhancer clusters in embryonic stem cells." eLife **3**: e04236.

Loh, Y. H., Q. Wu, J. L. Chew, V. B. Vega, W. Zhang, X. Chen, G. Bourque, J. George, B. Leong, J. Liu, K. Y. Wong, K. W. Sung, C. W. Lee, X. D. Zhao, K. P. Chiu, L.

Lipovich, V. A. Kuznetsov, P. Robson, L. W. Stanton, C. L. Wei, Y. Ruan, B. Lim and H. H. Ng (2006). "The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells." Nat Genet **38**(4): 431-440.

Luisi, B. F., W. X. Xu, Z. Otwinowski, L. P. Freedman, K. R. Yamamoto and P. B. Sigler (1991). "Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA." Nature **352**(6335): 497-505.

Manders, E., J. Stap, G. Brakenhoff, R. Van Driel and J. Aten (1992). "Dynamics of three-dimensional replication patterns during the S-phase, analysed by double labelling of DNA and confocal microscopy." Journal of cell science **103**(3): 857-862.

Martínez- Balbás, M. A., U. M. Bauer, S. J. Nielsen, A. Brehm and T. Kouzarides (2000). "Regulation of E2F1 activity by acetylation." The EMBO journal **19**(4): 662-671.

Maruyama, M., T. Ichisaka, M. Nakagawa and S. Yamanaka (2005). "Differential roles for Sox15 and Sox2 in transcriptional control in mouse embryonic stem cells." J Biol Chem **280**(26): 24371-24379.

Matsumura, I., H. Tanaka and Y. Kanakura (2003). "E2F1 and c-Myc in cell growth and death." CELL CYCLE-LANDES BIOSCIENCE- **2**(4): 333-338.

Mazza, D., A. Abernathy, N. Golob, T. Morisaki and J. G. McNally (2012). "A benchmark for chromatin binding measurements in live cells." Nucleic acids research: gks701.

Miki, T., S.-y. Yasuda and M. Kahn (2011). "Wnt/ β -catenin signaling in embryonic stem cell self-renewal and somatic cell reprogramming." Stem Cell Reviews and Reports **7**(4): 836-846.

Miller, J., A. D. McLachlan and A. Klug (1985). "Repetitive zinc-binding domains in the protein transcription factor IIIA from *Xenopus* oocytes." EMBO J **4**(6): 1609-1614.

Moseley, J. B. (2013). "An expanded view of the eukaryotic cytoskeleton." Molecular biology of the cell **24**(11): 1615-1618.

Mundra, P. A. and J. C. Rajapakse (2010). "SVM-RFE with MRMR filter for gene selection." NanoBioscience, IEEE Transactions on **9**(1): 31-37.

Murphy, K. P., Y. Weiss and M. I. Jordan (1999). Loopy belief propagation for approximate inference: An empirical study. Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc.

Nielsen, H., J. Engelbrecht, S. Brunak and G. V. Heijne (1997). "A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites." International journal of neural systems **8**(05n06): 581-599.

Niwa, H., Y. Toyooka, D. Shimosato, D. Strumpf, K. Takahashi, R. Yagi and J. Rossant (2005). "Interaction between Oct3/4 and Cdx2 determines trophectoderm differentiation." Cell **123**(5): 917-929.

Ogata, K., H. Hojo, S. Aimoto, T. Nakai, H. Nakamura, A. Sarai, S. Ishii and Y. Nishimura (1992). "Solution structure of a DNA-binding unit of Myb: a helix-turn-helix-related motif with conserved tryptophans forming a hydrophobic core." Proc Natl Acad Sci U S A **89**(14): 6428-6432.

Pabo, C. O. and M. Lewis (1982). "The operator-binding domain of lambda repressor: structure and DNA recognition." Nature **298**(5873): 443-447.

Pabo, C. O. and R. T. Sauer (1992). "Transcription factors: structural families and principles of DNA recognition." Annu Rev Biochem **61**: 1053-1095.

Palade, G. E. (1992). "Intracellular aspects of the process of protein secretion." Physiology Or Medicine: 1971-1980 **19711980**: 177.

Parraga, A., L. Bellolell, A. R. Ferre-D'Amare and S. K. Burley (1998). "Co-crystal structure of sterol regulatory element binding protein 1a at 2.3 Å resolution." Structure **6**(5): 661-672.

Pavletich, N. P. and C. O. Pabo (1991). "Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 Å." Science **252**(5007): 809-817.

Peng, T., G. M. Bonamy, E. Glory-Afshar, D. R. Rines, S. K. Chanda and R. F. Murphy (2010). "Determining the distribution of probes between different subcellular locations through automated unmixing of subcellular patterns." Proceedings of the National Academy of Sciences **107**(7): 2944-2949.

Pineda-Bautista, B. B. and J. Carrasco-Ochoa (2011). "General framework for class-specific feature selection." Expert Systems with Applications **38**(8): 10018-10024.

Pollard, T. D. and G. G. Borisy (2003). "Cellular motility driven by assembly and disassembly of actin filaments." Cell **112**(4): 453-465.

Qiu, W., X. Xiao, L. Wang and D. Gong (2013). "A Novel Pseudo Amino Acid Composition for Predicting Subcellular Location of Proteins." Journal of Computers **8**(3): 764-771.

Sarda, D., G. H. Chua, K.-B. Li and A. Krishnan (2005). "pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties." Bmc Bioinformatics **6**(1): 152.

Sauer, R. T., D. L. Smith and A. D. Johnson (1988). "Flexibility of the yeast alpha 2 repressor enables it to occupy the ends of its operator, leaving the center free." Genes Dev **2**(7): 807-816.

Singan, V. R., T. R. Jones, K. M. Curran and J. C. Simpson (2011). "Dual channel rank-based intensity weighting for quantitative co-localization of microscopy images." Bmc Bioinformatics **12**.

Takahashi, K. and S. Yamanaka (2006). "Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors." Cell **126**(4): 663-676.

Tang, S., T. Li, P. Cong, W. Xiong, Z. Wang and J. Sun (2013). "PlantLoc: an accurate web server for predicting plant protein subcellular localization by substantiality motif." Nucleic acids research.

Wagner, R. T., X. Xu, F. Yi, B. J. Merrill and A. J. Cooney (2010). "Canonical Wnt/ β -catenin regulation of liver receptor homolog- 1 mediates pluripotency gene expression." Stem Cells **28**(10): 1794-1804.

Wintjens, R. and M. Rooman (1996). "Structural classification of HTH DNA-binding domains and protein-DNA interaction modes." J Mol Biol **262**(2): 294-313.

Wu, C.-F. J. (1986). "Jackknife, bootstrap and other resampling methods in regression analysis." The Annals of Statistics **14**(4): 1261-1295.

Xu, X., L. Lu, P. He and L. Chen (2013). "Protein localization prediction using random walks on graphs." BMC Bioinformatics **14**(Suppl 8): S4.

Yu, W., H. K. Lee, S. Hariharan, W. Bu and S. Ahmed (2009). "Quantitative neurite outgrowth measurement based on image segmentation with topological dependence." Cytometry Part A **75**(4): 289-297.

Zhao, T., M. Velliste, M. V. Boland and R. F. Murphy (2005). "Object type recognition for automated analysis of protein subcellular location." Image Processing, IEEE Transactions on **14**(9): 1351-1359.

Zhu, S., P. Matsudaira, R. Welsch and J. C. Rajapakse (2010). Quantification of cytoskeletal protein localization from high-content images. Pattern Recognition in Bioinformatics, Springer: 289-300.