Curtin Medical School

Effect of Macromolecular Crowding on the Diffusion of RNAs and RNA-protein Complexes in the Translation Machinery

Vijay Phanindra Srikanth Kompella

0000-0002-9392-2375

This thesis is presented for the Degree of
Doctor of Philosophy
of the
University of Aberdeen
and
Curtin University

January 2022

Declaration

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgment has been made.

This thesis has been written as part of the collaborative doctoral programme undertaken at University of Aberdeen and Curtin University. This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

Vijay Phanindra Srikanth Kompella
Signature
Date: 29-Jan-22

Abstract

Aminoacyl-tRNA molecules deliver amino acids to the A-site of a translating ribosome. The availability of these tRNA molecules at the ribosome is crucial for the maintenance of optimal translation rates. The tRNAs that bind to the ribosome occur in ternary complexes with elongation factor proteins, and the diffusion of these protein complexes is affected by the cytoplasmic environment of cells. Cells are densely packed with macromolecules with 5-40% of their volume inaccessible for free diffusion, leading to excluded volume effects. Molecules inside cells also form short-lived clusters leading to further reduction of their diffusion. The extent to which these phenomena affect translation is not well understood. This study reports the use of a Brownian dynamics simulation approach to characterize the diffusion properties of tRNAs and ternary complexes in a model yeast cell environment. The protein composition of this environment was defined following a rigorous analysis of yeast proteomics datasets. The robustness of the Brownian dynamics approach suited for a study of this scale was investigated using an experimentally well-studied system comprising chymotrypsin inhibitor 2 as tracer, and bovine serum albumin or lysozyme as crowders, with a focus on characterization of slow- and sub-diffusion. The findings of this study indicate that, under normal cell-like conditions, the diffusion of tRNAs and ternary complexes is reduced by ~7-fold (compared with dilute conditions), whilst the diffusion under simulated severe osmotic stress conditions showed 70-fold decrease. The molecules also exhibited sub-diffusive behaviour which was stronger in the presence of osmotic stress. Investigations into the diffusion in crowded protein solutions revealed that cage effect causes sub-diffusion, and microsecond-scale slow diffusion is caused by excluded volume effects. The findings of this study can be readily used to accurately predict protein translation dynamics, including the crucial process of tRNA delivery to the ribosome, under a variety of conditions.

Attribution statement

As a part of my PhD, I co-authored a published research paper. The title of this paper is "Definition of the Minimal Contents for the Molecular Simulation of the Yeast Cytoplasm" (Kompella, V. P. S., Stansfield, I., Romano, M. C., & Mancera, R. L. (2019). Frontiers in Molecular Biosciences, 6, 97. https://doi.org/10.3389/fmolb.2019.00097). I am the first author of this paper. All the other authors contributed to the conception and design of this study. I conducted all the analysis and contributed to writing the manuscript, which was edited and proofread by Prof. M. Carmen Romano, Prof. Ian Stansfield, and Prof. Ricardo Mancera.

The contents of this paper are included in Chapter 4. Parts of the introduction of this paper are used in the introduction of Chapter 4. Results, methods, and discussion sections of the paper constitute the respective sections in Chapter 4.

I, Vijay Phanindra Srikanth Kompella, author of this document, declare that all the material that has been taken from the manuscripts described above was conceptualized and written by me.

Vijay Phanindra Srikanth Kompella

Signature

I, Ian Stansfield, as a co-author of the manuscript described above, endorse that the material used by the PhD candidate in this thesis is an accurate reflection of his contributions.

Ian Stansfield

Signature

I, M. Carmen Romano, as a co-author of the manuscript described above, endorse that the material used by the PhD candidate in this thesis is an accurate reflection of his contributions.

M. Carmen Romano

Signature

I, Ricardo. L. Mancera, as a co-author of the manuscript described above, endorse that the material used by the PhD candidate in this thesis is an accurate reflection of his contributions.

Ricardo. L. Mancera

Signature

Copyright statement

I have obtained permission from the copyright owners to use any third-party copyright material reproduced in the thesis, or to use any of my own published work in which the copyright is held by another party. The details supporting the same are presented in the appendix.

Funding source acknowledgement

This research project was supported by the scholarship provided under the Aberdeen-Curtin Alliance collaborative Ph.D. program, Curtin completion scholarship, and the extra funding generously provided by the Institute of medical sciences (Univ. of Aberdeen).

Acknowledgement of Country

We acknowledge that Curtin University works across hundreds of traditional lands and custodial groups in Australia, and with First Nations people around the globe. We wish to pay our deepest respects to their ancestors and members of their communities, past, present, and to their emerging leaders. Our passion and commitment to work with all Australians and peoples from across the world, including our First Nations peoples are at the core of the work we do, reflective of our institutions' values and commitment to our role as leaders in the Reconciliation space in Australia.

Acknowledgements

Countless people have directly or indirectly contributed towards my journey through my PhD, and I am extremely grateful to them. I wonder if I would have been able to even start a PhD without some of the support I received. I would like to start with my deceased father whose farsightedness and support allowed me to consider a PhD. I am greatly indebted to my mother who faced many challenges following my father's death but was brave enough to send me abroad in pursuit of my passion to do science. Despite her very traditional upbringing and her being not so well educated, she reinvented herself while I was away, and learnt new ways of living, especially during the COVID pandemic. Her sacrifices and support are one of the major reasons why this PhD was possible.

My supervisors at the University of Aberdeen, Ian (Prof. Ian Stansfield) and Mamen (Prof. Maria Carmen Romano), and my supervisor at Curtin University, Ricardo (Prof. Ricardo Mancera), have been of enormous support in both personal and professional spheres of my life as a graduate student. They have been critical of my work, provided timely feedback, helped me prioritise better, and navigated me through some of the toughest stages of the project. This complex interdisciplinary project would have been no less than a nightmare in the absence of their support and guidance. I am grateful for their assistance and efforts in securing the necessary supercomputing time in the UK and Australia. I am also indebted to them for leaving no stone unturned in finding the much-needed extra funding towards the end of the project. I benefitted enormously from their collective experience during the worst of the health and personal life crises that I endured, and I am very thankful to them for the same.

I would like to thank all the members of my defence committee (Prof. Rebecca Wade, Prof. Adrian Elcock, Prof. Valentina Tozzini, and Dr. Alessandro De Moura) for patiently going through my thesis and critiquing it. I am very grateful to Dr. Stefan Richter of the Heidelberg Institute for Theoretical Studies for his support with the SDA software. I am grateful for the supercomputing support I received from ARCHER (UK national supercomputing facility), Cirrus (UK National Tier-2 HPC Service at EPCC), and Pawsey supercomputing centre (Australia). I thank the University of Aberdeen and Curtin University for all the facilities and funding provided. I also thank the Institute for Complex Systems and Mathematical Biology (ICSMB) and Curtin Health Innovation Research Institute for the same.

I am grateful to my partner Anisha for her constant support through the toughest and busiest phases of the project and thesis writing. Without her support I doubt I would have had the mental energy to finish writing this thesis. I would like to express my deepest gratitude to Prof. P. V. Balaji at the Indian Institute of Technology-Bombay for his mentorship during my early days of science exploration. I thank my sister Ramya for being the best friend I could ask for during the worst and the happiest moments of my PhD journey. I could not have undertaken this journey without the emotional support and guidance I received from my aunt (whom I consider as my mother), Rajeswari. I am very grateful to my brother, Pavan, for doing all the housekeeping in India while I was away for my PhD. I'd also like to acknowledge Anisha's parents for the comforting conversations.

I have been away from home for nearly 15 years now, and friends have become my family. My colleagues, during my PhD, have become such good friends that there is no better place than here to acknowledge their contribution. I cannot thank enough John for his amazing company, insightful conversations, and for his patience to listen to my rants. I am very grateful to Chris for all the informative discussions, and for being there whenever I needed to think out loud or I needed someone to have a chat with. I thank Lanie for sharing her GROMACS wisdom and being the listener that I needed. I am very grateful to Sandra for listening to my stories, her insightful suggestions, and showing me the lighter side of life. I am also indebted to her for introducing me to their dog, Shadow, in whose company I forgot all the stress of the world and become a gentler soul. I am very thankful to Carlo for being there for me always. Thanks to the amazing company of all the other present and former BMMG members (Krushna, Kim, Lina, Yvonne and other honours students) my PhD journey did not feel too tedious. Many thanks to Arthur and Meng of the University of Aberdeen, in whose company the harsh Aberdeen winter was bearable. I also thank Rob, Sakshi, Deepali and countless others whom I might have missed naming here for being wonderful company in Perth.

Thanks to Teja (and his family), Pravin, Kranthi, Vikas, the 113 family, and all the other friends in India who were always a call away and managed to find time for me, whenever I needed them, despite their very busy routine. Lastly, I would like to thank the government of Western Australia for minimizing the impact of COVID, a crucial contribution towards my mental and academic wellbeing, for nearly 2 years. I am supremely confident that I must have missed acknowledging many others who contributed significantly towards this journey. To all those people, I am indebted to you for being there whenever I needed you.

Table of contents

Declaration ii
Abstract iii
Attribution statement v
Copyright statement vii
Funding source acknowledgement ix
Acknowledgement of Country xi
Acknowledgements xiii
Table of contents xv
List of Figures xvii
List of Tables xxi
List of equations xxiii
Chapter 1 Introduction 1
1.1 The molecular biology of gene expression 1
1.2 The role of transfer RNAs as carriers of amino acids 3
1.3 tRNA diffusion and channelling: the journey between the tRNA synthetase and the ribosome 5
1.4 Mathematical models of translation 8
1.5 Considering the role of tRNA diffusion in models of translation 10
1.6 Investigating the role of tRNA diffusion and channelling in translation using molecular dynamics simulation 11
1.7 Aims of the study 11
1.8 Structure of the thesis 12
Chapter 2 Literature review 13
2.1 Theory of diffusion and sub-diffusion 13
2.1.1 Continuous time random walk (CTRW) 15
2.1.2 Fractional Brownian motion (fBm) and fractional Langevin equation motion 15
2.1.3 Other models or combination of models of anomalous diffusion 17
2.2 Experimental approaches to characterizing diffusion under crowded conditions 18
2.2.1 FRAP 18
2.2.2 FCS 19
2.2.3 Pulsed field gradient NMR 20
2.2.4 Neutron backscattering 20
2.2.5 Single particle tracking (SPT) 21
2.2.6 Discussion 22
2.3 Computational approaches to characterizing diffusion under crowded conditions 24
2.3.1 Macromolecular crowding and diffusion 24
2.3.2 Cytosol like crowding and its effects 27
2.3.3 Discussion 33
Chapter 3 Characterization of slow and sub-diffusive behaviour in crowded protein solutions and discerning the underlying causal relations 35

3.1 Introduction 35
3.2 Approach and methods 36
3.2.1 Calculation of the α-exponent 37
3.2.2 Quantification of cage effects 38
3.3 Results 39
3.3.1 Diffusion coefficients and sub-diffusive behaviour 39
3.3.2 Cage effects in the protein crowded solutions 43
3.3.3 Non-Gaussianity and ergodicity 46
3.3.4 Excluded volume effects 51
3.4 Discussion and conclusions 56
Chapter 4 Definition of the minimal contents for the molecular simulation of the yeast cytoplasm 59
4.1 Introduction 59
4.2 Methods 60
4.2.1 Definition of a eukaryote cell simulation environment 60
4.2.2 Statistical analysis 60
4.3 Results 61
4.3.1 Analysis of internal consistency of yeast proteomics datasets 61
4.3.2 Selection of datasets 66
4.3.3 Constraints for the definition of the contents of a simulation cell 67
4.3.4 Definition of the contents of the simulation cell 69
4.4 Discussion 70
Chapter 5 Characterization of the diffusion properties of tRNAs and their complexes in the model yeast cytoplasm 79
5.1 Introduction 79
5.2 Methods 80
5.2.1 Pre-processing of Ribosome 80
5.2.2 Pre-processing of tRNA and ternary complex 81
5.2.3 Pre-processing of Protein crowders 82
5.2.4 Determination of the composition of tRNAs 83
5.2.5 Polydispersity index (PDI) 83
5.2.6 Simulation system 84
5.3 Results 86
5.3.1 Slow diffusion of tRNAs in the crowded cytoplasm 86
5.3.2 Sub-diffusion of tRNA and the EF-1α ternary complex 89
5.4 Conclusions 91
Chapter 6 Conclusions and future directions 95
References 99
Appendix 119
Content permission 119
Figure permissions 119

List of Figures

Figure 1.1 Schematic representation of transcription and translation processes. 1

Figure 1.2 Schematic representation of different steps in the translation process. It should be noted that the start/stop codons vary depending on the organism. 3

Figure 1.3 Schematic representing the channelling process postulated in eukaryotes. 6

Figure 1.4 Density(rho) of the particles calculated using numerical simulations at different entry (α) and exit (β) rates. The thick lines represent boundaries calculated from the mean-field approach. 9

Figure 2.1 The displacement distributions of the simulations conducted using fBm in a finite interval under (a) sub-diffusive, (b)normal diffusive and (c) super-diffusive conditions. A clear deviation from Gaussian distribution is observed for long time simulations. The solid lines represent ideal Gaussian distributions at corresponding times. This figure was taken from the work of Guggenberger et al.[64] (doi: https://doi.org/10.1088/1367-2630/ab075f) 17

Figure 2.2 The trajectories of tRNAs in E.coli cells revealed from the tracking of fluorescently labelled tRNAs is shown in palette 'a' and palette 'c' indicates that the distribution of apparent diffusion coefficient arises from the diffusion properties of two species, the red line corresponding to ribosome bound slow species and blue line corresponds to fast-diffusing, presumably unbound tRNAs. This figure is taken from the work of Plochowietz et al.[30] (doi: https://doi.org/10.1093/nar/gkw787) 22

Figure 2.3 Evolution of the detail at which model cytoplasm is represented over the course of years. (i) Cytoplasm model of Ridgway et al.[25] at three different volume fractions represented by a, b and c. (Appendix: figure permissions 1). (ii) Hasnain et al.'s cytoplasm model[121] with macromolecules represented as spheres or cluster of spheres. (doi: https://doi.org/10.1371/journal.pone.0106466) (iii) Palette B cropped out of figure 1 in Yu et al.'s paper.[114] The figure shows macromolecules represented at the atomistic level of description with explicit solvent. (doi: https://doi.org/10.7554/eLife.19274.001) 30

Figure 2.4 Palette 'a' of figure 6 in Trovato et al.'s paper.[108] The figure shows the variation of α-exponent and ergodicity breaking parameter(EB) with respect to the size of the crowder. The system is ergodic when EB is 0 and is non-ergodic when EB=1. A clear correlation between increase in non-ergodicity and sub-diffusion can be seen. The green colored lines correspond to simulations with attractive and repulsive interactions, whereas the black colored lines show the data from simulations with repulsive-only interactions. The three regions in the plot labelled as region I, II, and III correspond to particles of different sizes. Region II corresponds to particles of the size of ribosomes, region I and III correspond to particles that are smaller (most of the protein molecules) and larger than ribosomes (nucleiod) respectively. (Appendix: figure permissions 2) 31

Figure 3.1 Quantification of cage effects. The particle is represented in yellow. Arbitrary vector perpendicular to r_{01} is represented as a dotted vector. 39

Figure 3.2 Predicted diffusion properties in crowded protein solutions. (A) Comparison of experimental and predicted CI2 diffusion coefficients. The predicted values are within the same order of magnitude of experiment, revealing good agreement. (B) The predicted and observed diffusion coefficients of the CI2 tracer in the presence of the protein crowders BSA and lysozyme, and of the protein crowders themselves are plotted as a function of crowder concentration. As expected, the increase in crowder concentration results in a downward trend of the diffusion coefficient of CI2. (C) Average x12 as a function of |r01| (green), whilst the dashed red line corresponds to the reference x = 0 curve, and the dotted vertical line separates the regions of low and high noise. The yellow line corresponds to the linear fit for the less noisy region, whose slope is used in the calculation of α-exponent. The slope is negative, indicating the presence of caging effects. (D) Average y12 as as a function of |r01| (green), whilst the dashed dotted line corresponds to the reference y = 0 curve, and the blue dotted line (which is very close to the y = 0 curve) corresponds to the linear fit of the less noisy region. |r01|, x12 and y12 are all provided in Å. Plots C and D correspond to data at a BSA concentration of 300 g/L at 5ns. 41

Figure 3.3. Sub-diffusive and non-Gaussianity properties of the crowders and tracer (50 and 300g/L concentration of crowder). The blue, orange and grey lines in all the curves represent the α-exponent calculated from the log(TAMSD/τ) vs log(τ) curves, α-exponent calculated from cage effects, and non-Gaussianity parameter (NGP) measured at different lag times respectively. All the curves on the left side of the figure represent the data for low concentration of the crowder at 50g/L and the ones on the right side represent data for high crowder concentration. The data for CI2 in BSA is in the first row highlighted in green, followed by data for CI2 in LYS in next row highlighted in yellow, followed by data for LYS and BSA highlighted in orange and red respectively. Error bars represent the standard deviation of the value of the α-exponent between simulations started with different configurations. The time ranges in the individual graphs are different from each other due to the variation in the emergence of noise in the log(TAMSD/τ) vs log(τ) curves. 42

Figure 3.4. Sub-diffusive and non-Gaussianity properties of the crowders and tracer (at concentrations of 100 and 200 g/L of the crowder). The data is represented in the same way as in Figure 3.3. 43

Figure 3.5. Variation of the intensity of cage effects with respect to time and crowder concentration. The straight lines plotted are representative of the slope calculated from the less noisy regions of plots of <x12> or <y12> vs |r01|. The blue, red and yellow lines represent slopes at short, intermediate and long time scales, respectively. The first and second rows highlighted in green and yellow represent the data for the diffusion of CI2 in BSA and lysozyme, respectively. The next two rows highlighted in orange and red represent the data for the self-diffusion of lysozyme and BSA, respectively. The first two columns of every row contain plots of <x12> vs |r01| and <y12> vs |r01| (in that order) at the low protein crowder concentration of 50 g/L. The last two columns contain the same plots at the high protein crowder concentration of 300 g/L. 45

Figure 3.6. Convergence of the diffusion coefficient of BSA at different concentrations in simulations with the full energy term. The error bars represent the standard deviation (n=3). 48

Figure 3.7. Convergence of the diffusion coefficient of BSA at different concentrations in simulations with the soft-core repulsive term only. The error bars represent the standard deviation (n=3). 49

Figure 3.8. Convergence of the diffusion coefficient of lysozyme at different concentrations in simulations with the full energy term. The error bars represent the standard deviation (n=3). 49

Figure 3.9. Convergence of the diffusion coefficient of lysozyme at different concentrations in simulations with the soft-core repulsive term only. The error bars represent the standard deviation (n=3). 50

Figure 3.10 The properties of tracer and crowder in the absence of attractive interactions (at concentrations of 50 and 300 g/L of the crowder).The data is represented in the same way as in Figure 3.3. The value of the α-exponent calculated using the log plot and cage effect, and NGP are computed for systems without attractive interactions. 52

Figure 3.11 The properties of tracer and crowder in the absence of attractive interactions (100 and 200 g/L of the crowder). Error bars represent standard deviation (n=3). The data is represented in the same way as in Figure 3.10. 53

Figure 3.12. Radial distribution functions of BSA and LYS. The red curves correspond to simulations with the soft-core repulsive term only and the blue curves correspond to simulations with the full energy term. The effective radius was approximated as the maximum distance (r) at which RDF~0, and it increases in value by 2.8 Å and 1.6 Å in BSA and LYS respectively, when only the soft-core repulsive term is used. Due to the larger size of BSA compared with LYS, the change in excluded volume due to the small change in the effective radius is more pronounced in the former. 56

Figure 4.1. Distribution of protein mass (calculated as the product of molecular weight times abundance) per cell plotted as a function of the mass rank of each protein. Proteins in the yeast proteomics dataset were ranked according to their mass, exhibiting a clear exponential decrease as a function of their mass rank in the cell. In the inset the cumulative percentage of mass is plotted as a function of rank. The top 200 cytoplasmic proteins contribute to approximately 70% of the total cell protein mass. 62

Figure 4.2. Statistical analyses of proteomics datasets. (A) Pairwise correlations between the ontological profiles obtained for the individual datasets. Correlations were measured using the Pearson correlation coefficient, whose values are colour-coded (from the highest correlation in yellow to the lowest correlation in

blue). (B) The ontology profile overlap between datasets is quantified using the Jaccard index and the colour-code is the same as in the previous panel. In both panels mass spectrometry based datasets are indicated in red on the axes labelled as LU [150], PENG [151], KUL [152], LAW [153], LAHT [154], DGD [155], PIC [156], LEE2 [157], THAK [158], NAG [159] and WEB [160]; GFP datasets are shown in green on the axes and are labelled as TKA [161], BRE [162], DEN [163], MAZ [164], CHO [165], YOF [149], NEW [166], LEE [167] and DAV [168]; and the TAP-immunoblot dataset is shown in white on the axes and is labelled as GHA [169]. The top 200 proteins are shown to have a similar gene ontology profile across all of the datasets. 63

Figure 4.3. Testing of statistical difference between the abundance of ribosomal proteins in each of the datasets. Mass spectrometry-based datasets are shown in red on the axes, GFP datasets are shown in green on the axes and the TAP-immunoblot dataset is shown in white. Ribosomal protein numbers were not reported in the YOF dataset and, therefore, it is not included. The results of t-tests with $p > (0.05/190)$ are coloured dark blue and all others are coloured light blue. GFP datasets exhibit a high level of consistency. There is also consistency among the first five MS datasets. However, there are no discernible patterns in terms of the growth media, growth phase or protein abundance units. 65

Figure 4.4 Results of the t-tests corrected for type-I errors using the Benjamini-Hochberg approach (an alternative to the Bonferroni correction) with FDR=0.05. Dataset pairs for which p-values > 0.05 are colored in dark blue. Squares in red show deviations from the t-test predictions. The results are qualitatively similar to the t-test predictions and the conclusions drawn from t-tests remain valid. 71

Figure 4.5. Results of the Mann-Whitney U test performed in pairwise manner across the datasets. The Bonferroni correction was applied to address type-I errors. Squares in dark blue show p-values > (0.05/190). Squares in red show the dataset pairs for which the p-values predicted using the Mann-Whitney U test are different from the p-values predicted with the t-tests. The results are qualitatively similar to the t-test results and the conclusions drawn from the t-tests remain valid..... 72

Figure 5.1. Snapshots of the simulation systems of the crowded systems. The RNA molecules in all the systems are shown in red. Red lines show the boundaries of the simulation cell. (A) Whole cytoplasm (WHOLE). The ribosome is the largest molecule in this system, with its proteins shown in yellow and RNA in red. tRNAs are shown in red, elongation factor proteins are shown in blue, and the other crowding proteins are shown in random colours. The edge length of the cubic simulation cell is 560 Å. (B) REDUCED system, where the top four most abundant proteins are the crowders. The macromolecular density of this system is 90 g/L with a simulation cell edge length of 560 Å. Based on the visibly less empty space between the crowders, it can be inferred that the macromolecular density corresponding to the ribosome is distributed evenly across the simulation box. (C) REDUCED HIGH system. The edge length of the cubic simulation cell is 388 Å, with a macromolecular concentration of 270 g/L. Due to the close packing of proteins and the two-dimensional representation of the box, the spaces between molecules are not easily discernible but even distribution of proteins can be assumed. 86

Figure 5.2. Convergence of the predicted diffusion coefficients of the ternary complex of EF-1α with tRNA in the (A) whole cytoplasm (no.of molecules = 13), (B) reduced cytoplasm with high concentration of macromolecules (no.of molecules = 12), and (C) reduced cytoplasm with normal concentration of macromolecules (no.of molecules = 13). 87

Figure 5.3. Plots of log (MSD/t) vs log (t). The blue-, red- and green-coloured lines represent the data of tRNAC, tRNAUC and tRNAEF, respectively from the three initial configurations of a given system. The ranges between X-axis values 0 and 2, 2 and 3 represented by the black vertical lines show the ranges for which the α-exponent is calculated. 90

List of Tables

Table 3.1 Ratios of the long-time diffusion coefficients measured in simulations with only soft-core repulsive interactions in simulations with the full energy term. Dark green colours indicate a high ratio..... 55

Table 4.1. Mean abundances (averaged over 21 datasets) of ribosomal proteins. 72

Table 4.2 The final list of proteins and their structural information. The first column is the rank of the protein when the list is sorted in descending order of mass contributed to the simulation cell. Rows are colour-coded such that green denotes proteins that have an experimentally-determined structure (completely or partially), white denotes proteins that do not have structures but the structures can be predicted using homology modelling, and yellow denotes proteins that do not show sequence similarity to any known structure. There are structures readily available for 34 of the protein types, whilst 32 of the protein types show significant sequence identity with protein structures available and, therefore, their structures can readily be obtained using homology modelling. The remaining 4 types of proteins show no sequence similarity to any structures publicly available and, therefore, *ab initio* modelling approaches can be used to predict their structures. 74

Table 5.1 Simulation cell contents of the different types of simulations. Charged tRNAs, uncharged tRNAs and ternary complexes are represented by tRNAC, tRNAUC, and tRNAEF respectively. 85

Table 5.2. Predicted diffusion coefficients of tRNA molecules. Standard deviations (n=3) of the data from three initial configurations are shown in the brackets..... 89

Table 5.3. Predicted alpha exponents of tRNA molecules. Standard deviation (n=3) corresponding to the data from three initital configurations is shown in the brackets. 90

List of equations

Equation 2.1 13
Equation 2.2 14
Equation 2.3 14
Equation 2.4 15
Equation 2.5 16
Equation 2.6 16
Equation 2.7 16
Equation 2.8 19
Equation 2.9 19
Equation 2.10 20
Equation 2.11 27
Equation 2.12 27
Equation 2.13 31
Equation 3.1 38
Equation 3.2 46
Equation 3.3 47
Equation 4.1 68
Equation 4.2 68
Equation 4.3 68
Equation 5.1 83

Chapter 1 Introduction

1.1 The molecular biology of gene expression

The information necessary for survival and reproduction of living organisms is encoded in the sequence of nucleotides in the DNA. It is decoded as the sequence of amino acids in proteins, that govern the functioning of the cell in a process mediated by RNA. The first step in this process is termed 'transcription', in which the sequence information is copied onto an RNA molecule. RNA thus synthesized undergoes post-transcriptional modifications to produce messenger RNA (mRNA). Ribosomes translate the information in the mRNA in the presence of aminoacyl transfer RNAs (aa-tRNAs), in a process called 'translation', leading to the production of proteins. (Figure 1.1) A gene is the sequence of nucleotides in the DNA upon which RNA and eventually proteins are synthesized, and the process of formation of these molecules from a gene is called 'gene expression'.[1] Gene expression can therefore be regulated at the level of transcription and/or translation. The mechanism of regulation of translation can be understood by inspecting more closely its biochemistry.

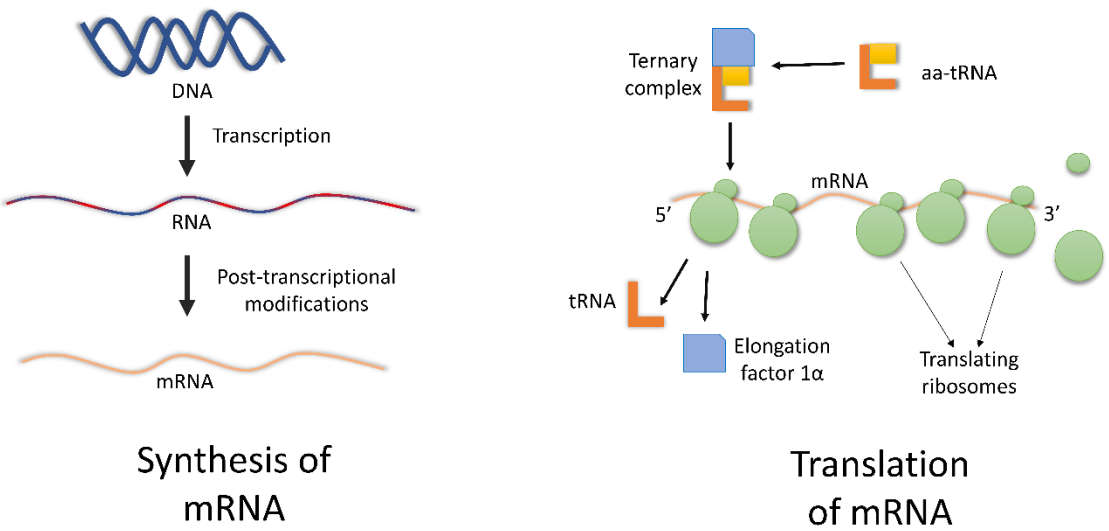

Figure 1.1 Schematic representation of transcription and translation processes.

Translation is divided into three steps: initiation, elongation and termination. During the initiation step, a complex of the small ribosomal subunit, initiation factors and the initiator tRNA binds to the start codon on the mRNA located near its 5' end. The large ribosomal subunit

then binds to this complex, releasing the initiation factors. There are three tRNA-binding sites on a ribosome, referred to as E, P, and A sites. An aminoacyl tRNA (aa-tRNA) transports the amino acid to the A-site of the ribosome, where it binds through its anti-codon to the mRNA codon at the A-site (acceptor site). The polypeptide-carrying tRNA is accommodated at the P-site (peptidyl site) and during the elongation step the polypeptide is transferred to the aminoacyl tRNA at the A-site through the formation of a peptide bond between the new amino acid and the polypeptide chain. Following this, the ribosome 'hops' to the next codon on the mRNA and, as a result, the new codon now occupies the A-site. The P-site is occupied by the tRNA with the polypeptide chain, and the tRNA that just transferred the polypeptide chain leaves the ribosome from the E-site (exit site). Following their exit, tRNAs are covalently bonded to the corresponding amino acids in a reaction catalysed by aminoacyl tRNA synthetases, resulting in the formation of aa-tRNAs. The elongation step takes place in a recursive manner until the ribosome encounters a stop codon, following which the ribosome exits the mRNA at the 3' end. (Figure 1.2)

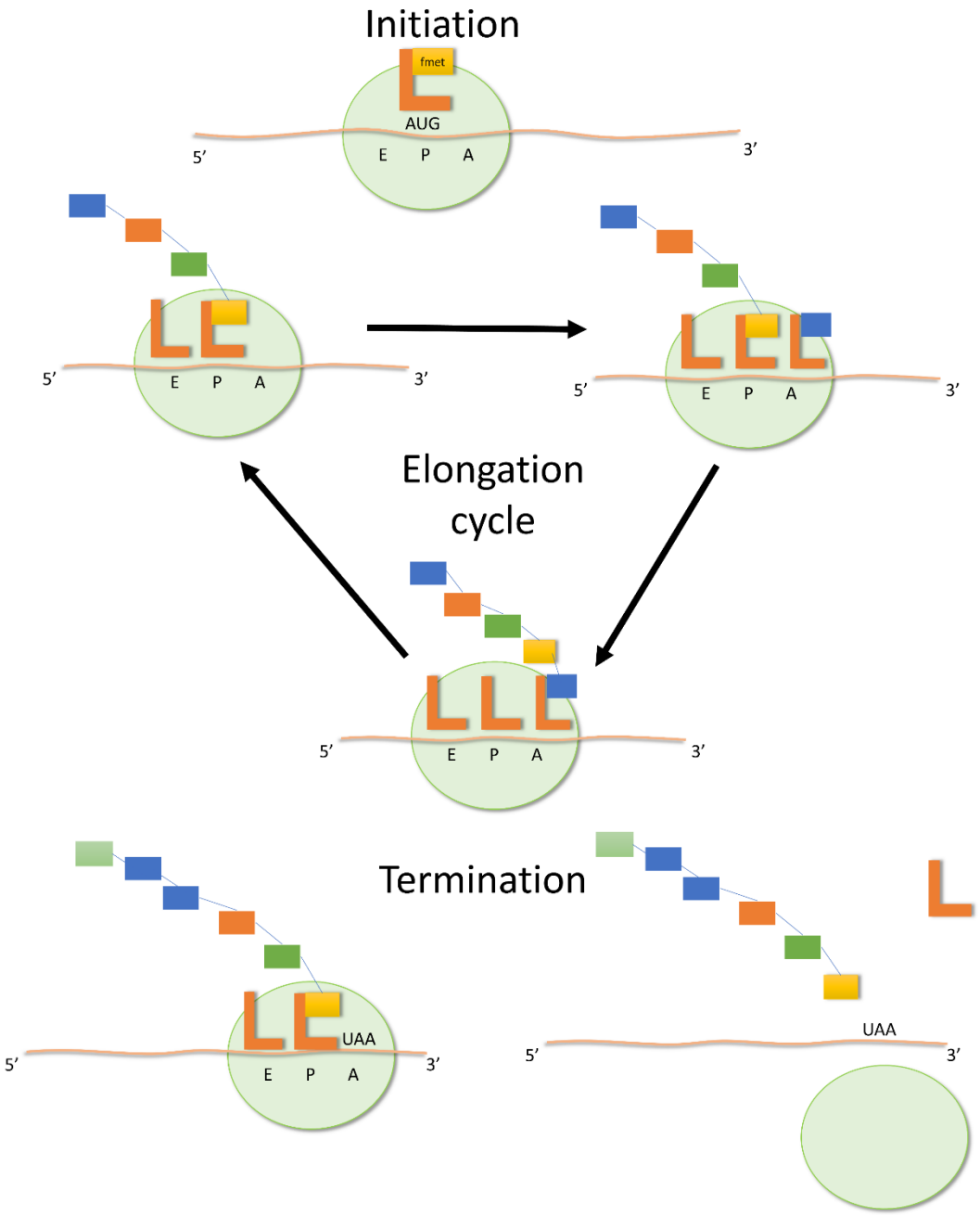

Figure 1.2 Schematic representation of different steps in the translation process. It should be noted that the start/stop codons vary depending on the organism.

1.2 The role of transfer RNAs as carriers of amino acids

Transfer RNAs (tRNAs) play a crucial role in the elongation step of protein translation. A total of 64 types of triplet codons are present in mRNA, 61 of which code for amino acids and the rest act as stop codons. The 20 amino acids present in the cells are coded by these 61 codons. Unconventional (wobble) base pairing allows the 61 codons to be read by fewer tRNAs, which

amounts to 41 in the case of *Saccharomyces cerevisiae*.[2] This results in different species of tRNAs carrying the same amino acid, and such tRNAs are termed isoacceptors.

The concentration of different species of tRNAs in the cytoplasm is not the same.[2] The rate at which a codon in mRNA is decoded is a function of the concentration of corresponding tRNAs.[3,4] The concentration of tRNAs depends on the supply of tRNAs, maintained by their gene expression, and the demand for them, regulated by the codon composition of mRNAs present in the cytoplasm at that moment in time.[2] In the case of 'rare codons', the concentration of required tRNAs is low, which causes ribosomal pausing or drop off.[5,6] Rare codons can be replaced with synonymous codons, which translate to the same amino acid but make use of isoacceptors that are more readily available to optimize translation rate. Such 'codon optimization' has applications in gene therapy and nucleic acid based vaccines.[7] Highly expressed genes in a functional class were shown to have fewer rare codons in a genome level study of protein translation.[8] It was also observed that polyglutamine-associated mRNAs, which repeatedly call for a single species of tRNA and deplete the corresponding tRNA, are associated with Huntington's disease.[9] The timely supply of aa-tRNAs is thus a crucial aspect of translation.

The aa-tRNAs reach the A-site in the form of a ternary complex which consists of translational elongation factor (EF1α), aa-tRNA, and GTP. Three types of such complexes arrive at the A-site classified based on the degree of codon-anticodon complementarity; (i) complexes with aa-tRNAs showing complementarity (including wobble pairing) termed 'cognate complexes', (ii) complexes with a lower degree of complementarity called 'near-cognate complexes', (iii) complexes with no complementarity termed 'non-cognate complexes'. Ribosomes read and filter these complexes with a high degree of accuracy at an error rate of 10^{-5} to 10^{-3} across prokaryotes and eukaryotes.[10] The kinetics of the steps between the binding of ternary complexes and the formation of peptide bond regulate the error rates. More specifically, the hydrolysis of GTP by the elongation factor (in the ternary complex) and the accommodation of aa-tRNA following this hydrolysis exhibit widely different kinetics for cognate and near-cognate complexes.[11] Mismatches at a level of single base-pair result in a 1000-fold increase in the rate of disassociation of ternary complexes.[11] Although the error rates are low, due to the sheer number of tRNA reading events occurring in a cell, there is at least one mistranslated amino acid in 15% of the average-sized proteins. Erroneously translated proteins exhibit

misfolding and may lead to cell death.[12] Therefore, availability of cognate ternary complexes in adequate numbers at the A-site is crucial.

1.3 tRNA diffusion and channelling: the journey between the tRNA synthetase and the ribosome

Once the aa-tRNA is delivered, the complex of EF1α and GDP exits the ribosomal A-site. The tRNAs exiting the ribosome, following the delivery of an amino acid, need to reach the synthetases to undergo aminoacylation. In a cascade of reactions executed by multiple enzymes (or multiple reaction centres in a single molecule), the transport of the intermediates between the reaction centres often occurs in a controlled manner. This is in contrast to the expected free diffusion of intermediates into bulk environment. Such controlled transfer of intermediates between the enzymes is termed 'channelling'. Intramolecular channelling is known to occur via electrostatic guidance in Krebs cycle[13] and through intramolecular tunnels during the transport of ammonia, aldehydes, and carbamates[14]. Spatial proximity of the reacting enzymes is also shown to increase reaction rates.[15] This occurs when the products of the first reaction do not diffuse fast (compared with their production rate) leading to increase in their local concentration, and the presence of the second enzyme in the proximity leads to higher reaction rates.[14] This type of channelling is controlled by the diffusion rates and the environment of diffusing particle. A similar channelling mechanism has been postulated to explain the delivery of ternary complexes to the A-site, delivery of uncharged tRNAs to the synthetases and the subsequent formation of ternary complexes.

A series of experiments conducted by Deutscher and co-workers provided evidence for 'tRNA channelling' in mammalian cells. In the first set of experiments, ^{3}H-labelled aa-tRNAs were transfected into permeabilized CHO cells along with free ^{14}C-labelled amino acids and, due to significantly low incorporation of ^{3}H in the proteins compared with ^{14}C, it was concluded that the aa-tRNAs are transferred to the ribosome through a channelling mechanism.[16] The second set of experiments indicated that the exogenous uncharged tRNAs do not enter the tRNA channelling cycle, which was inferred based on their inability to affect protein synthesis.[17] In both sets of experiments protein synthesis was observed for ~20 min. However, recent single cell level observations at time scales of ~7h following transfection and employing fluorescent labelled uncharged tRNAs transfected into CHO cells, indicated that the exogenous tRNAs co-localize with the translation machinery and participate in the translation process.[18] In a

sequence analysis study it is found that the consecutive occurrence of an amino acid is coded to use the same tRNA repeatedly in a phenomenon termed 'codon auto-correlation'.[19] Although tRNA channelling can be inferred from codon auto-correlation in *Saccharomyces cerevisiae*[19], there is no direct evidence for channelling in lower eukaryotes. Although there is evidence for channelling in higher eukaryotes, similar detailed data is not available for lower eukaryotes. Secondly, the more recent experiments on the CHO cells, as described above, indicate that exogenous tRNAs can enter the elongation cycle hinting at the possibility of diffusion of tRNAs into bulk. The simple dichotomy that arises here is the presence or absence of channelling. In the presence of channelling, the enzymes involved are in the proximity of each other, and slower diffusion of molecules ensures that the intermediates reach the target enzymes before leaking into the bulk. Therefore, diffusion of tRNAs and ternary complexes plays a crucial role in the manifestation of channelling. In the other scenario, free diffusion of these molecules occurs resulting in a similar important role for the diffusion of tRNAs and their complexes. (Figure 1.3)

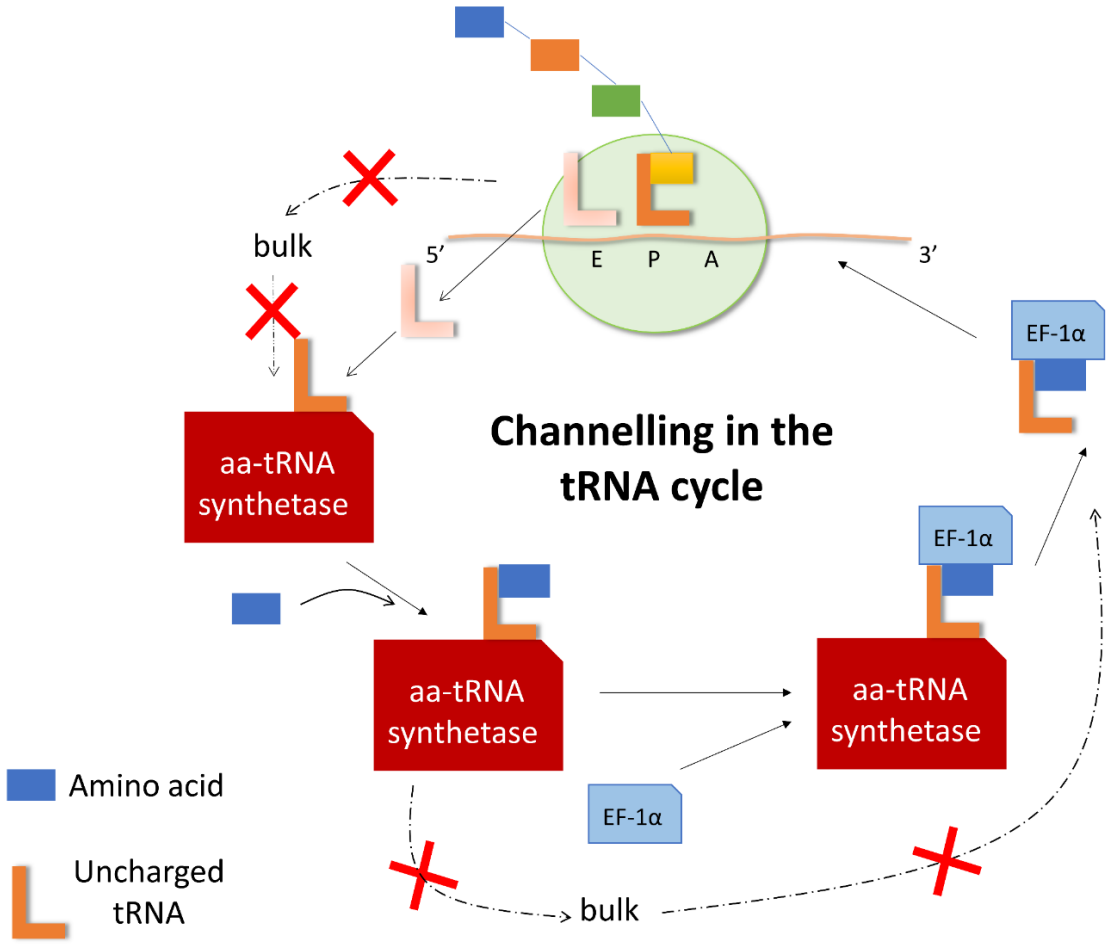

Figure 1.3 Schematic representing the channelling process postulated in eukaryotes.

Diffusion *in vivo* is different from that in dilute conditions. Cells are densely packed with macromolecules, typically at a concentration of 100-450g/L. About 5-40% of the cellular volume is occupied by macromolecules, leading to so-called excluded volume effects.[20] The extent of 'macromolecular crowding' depends on the cell type, its volume, growth rate, and differentiation stage.[21] As a result of excluded volume effects, the diffusion properties of metabolites and macromolecules are significantly affected. Molecular simulation studies on crowded protein solutions have shown that the solvent is mostly confined to first and second solvation layers.[22] It is believed that this leads to an altered dielectric response in water[22], which in turn affects molecular interactions. Such change in dielectric constant has been observed experimentally in yeast cells.[23] Excluded volume effects and altered interactions as a consequence of macromolecular crowding affect reaction processes in cells. For example, the binding and release of the substrate of alcohol dehydrogenase is shown to be affected by crowding.[24] The effect of crowding on the diffusion properties of barnase and barstar explains their association.[25] The diffusion coefficient of green fluorescent protein (GFP) has been reported to be a magnitude lower *in vivo* compared with that in an infinitely dilute system.[26] Multiple molecular simulation and experimental studies have shown that diffusion properties are significantly affected, leading to slow- and sub-diffusion due to crowding. This will be discussed in detail in Chapter 2.

Depending on its type, there are 76-90 nucleotides in a tRNA molecule.[27] The Stokes radius of the ternary complex calculated using HYDROPRO[28] is nearly 28 Å. In HYDROPRO, the surface of the molecule is represented as a shell of beads and the diffusion properties are calculated, in the presence of a medium with appropriate viscosity, using this simplified system.[29] Given these properties of the ternary complex and tRNAs, crowding is expected to significantly affect their diffusion in cells. The *in vivo* diffusion coefficient of tRNA has been measured using fluorescent labelled tRNAs in *E.coli*. The position of a single tRNA particle was captured every 5 ms over a period of 1.5 s, based on which the diffusion coefficient of free tRNAs (unbound to elongation factors or ribosomes) was measured to be 8.1 $\mu m^2/s$.[30] This is nearly ten times slower than the diffusion coefficients calculated using HYDROPRO under dilute conditions.[31] Due to the timescales explored, these experiments do not capture key characteristics, such as the sub-diffusion typically observed in crowded protein solutions at sub-microsecond scales.[32,33] The kinetic properties of translation in *E. coli*, derived assuming diffusion-controlled binding of ternary complexes to the ribosome, match closely to those obtained from experimental data.[34] However, a mechanistic approach detailing the role of

crowding in a more specific manner is necessary to characterize these effects quantitatively. In eukaryotes, however, these effects have not been characterized either at a mechanistic or phenomenological level of detail.

1.4 Mathematical models of translation

Several mathematical models have been developed to explain translation dynamics in a quantitative manner.[2,8,35–40] Translation is often modelled as a totally asymmetric simple exclusion process (TASEP), in which mRNA is treated as a unidimensional lattice with the lattice sites corresponding to codons. The ribosomes enter the lattice at a certain rate, move stochastically along the lattice during the elongation process at a certain hopping rate (k), before exiting the lattice. The particles can only enter a lattice site that is unoccupied, and the particles cannot overtake each other during their unidirectional movement across the lattice. In the mean-field approach, which is an approximation, the states of individual lattice sites, represented by the presence or absence of particles in them, are assumed to be uncorrelated. Gillespie's stochastic simulation algorithm[41], is often used to simulate this process numerically.[42] The conclusions of numerical simulations and mean-field approach indicate the presence of phases that depend on the rates of entry(α), exit(β), and hopping (k) of the particles. The phase of the system is defined by its characteristic current and density of the particles on the lattice. The rate of protein synthesis is equivalent to the current on the lattice. There are four phases identified in a simple TASEP model, and assuming the hopping rate (k) to be constant, the following conclusions can be drawn. In the low-density (LD) phase the entry rate is low ($\alpha < k/2$) and the exit rate is greater than the entry rate ($\alpha < \beta$). The low entry rate and higher exit rate lead to low density (average density=α/k) of particles on the lattice. The current in this phase is $\alpha(1-\alpha/k)$. In high-density (HD) phase, exit rate is low ($\beta<k/2$) and the exit rate is lower than the entry rate ($\beta< \alpha$) leading to high particle density on the lattice. The average density and current in this phase are $(1-\beta/k)$ and $\beta(1-\beta/k)$ respectively. Maximal current (MC) phase is observed for $\alpha, \beta> k/2$, where the density of the lattice is 0.5. Finally, shock phase is observed when $\alpha=\beta$ and $\alpha, \beta< k/2$. (Figure 1.4). The results of the mean-field approximation can be shown to be exact in the limit of an infinite lattice.[43]

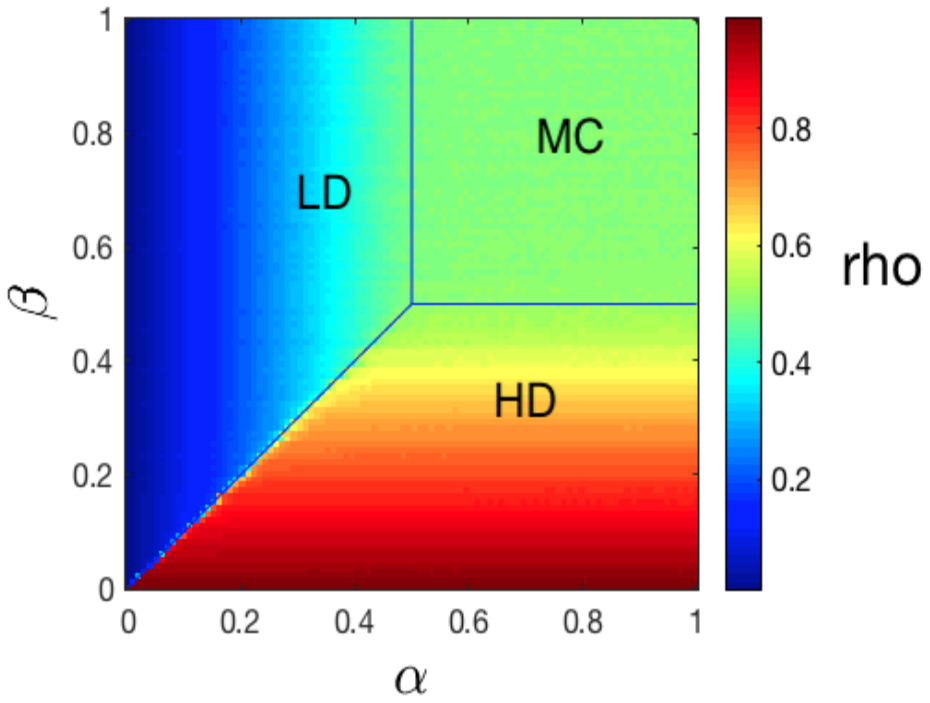

Figure 1.4 Density(rho) of the particles calculated using numerical simulations at different entry (α) and exit (β) rates. The thick lines represent boundaries calculated from the mean-field approach.

This simple TASEP model can be improved further to represent the translation process more accurately. For example, ribosomes wait for aa-tRNAs upon reaching A-site, and this waiting time is accounted for in a two-state model.[38] The model has been further improved by treating synthetase activity explicitly and its consequences on the supply of aa-tRNAs.[2] The translating ribosomes can also erroneously drop-off before reaching the non-sense codon towards the 3' end of the mRNA.[44] Some mathematical modelling approaches also account for such drop-off events while estimating the translation rates.[36,40,44] Ribosomes exiting the mRNA re-enter it by binding at the 5' end in a process termed 'ribosome recycling', which is often seen in eukaryotes.[44] Such recycling of ribosomes was also included in the modelling approaches.[44,45] The dynamics of simultaneous translation of multiple lattices competing for shared resources have also been characterized.[42]

1.5 Considering the role of tRNA diffusion in models of translation

Despite these improvements in the mathematical description of translation, the crucial role played by macromolecular crowding in regulating translation is not captured in these approaches. In an approach developed for yeast by Shah et al., tRNA and ribosomal diffusion are accounted for in an implicit manner while calculating the probability of initiation and elongation.[46] The diffusion coefficients[31] used in this model were computed for dilute conditions, hence not accounting for crowding effects. The ternary complexes are treated equivalently to free tRNAs and ternary complex formation dynamics are not treated explicitly.[46] The model proposed for *E. coli* by Zhang et al. is by far the best approach to describe the role of crowding on the diffusion of ternary complexes and the subsequent effects on translation. In their stochastic simulations, diffusion coefficients estimated based on molecular weight under crowded conditions are used to simulate the Brownian random walk of ternary complexes. Use of this approach indicated that local depletion of ternary complexes, caused by repeated request for a particular tRNA molecule, is a function of their diffusion coefficient, which in itself is affected by crowding. The global depletion induced by crowding, through the restricted diffusion of ternary complexes, was also shown to have a significant effect on translation rates.[40] However, since crowding effects are not treated explicitly, sub-diffusion observed in *E.coli*[47] is unaccounted for in these simulations.

In summary, the mathematical modelling approaches developed so far for translation have not accounted properly for the effects of macromolecular crowding in eukaryotes. Some of the models proposed for bacteria underline the importance of crowding in regulating translation dynamics. However, it is important to note that these models do not take into account the true complexity of the problem and fail to capture sub-diffusion potentially arising from crowding. A comprehensive understanding of the diffusion properties of ternary complexes and tRNAs is an essential first step to building mathematical models that account for macromolecular crowding effects on translation dynamics.

1.6 Investigating the role of tRNA diffusion and channelling in translation using molecular dynamics simulation

There are no *in vivo*, *in vitro* or *in silico* studies that have characterized the diffusion properties of eukaryotic ternary complexes or tRNAs taking into consideration the macromolecular crowding effect. tRNAs inside cells exist as free molecules or bound as ternary complexes or bound to ribosomes, and as a result of this it is not easy to experimentally characterize the diffusion properties of tRNAs at the resolution of individual complexes. Moreover, key properties like sub-diffusion, usually observed at sub-microsecond timescales, cannot be easily captured in experiments. A computational approach employing molecular dynamics (MD) simulation can capture crowding effects on tRNA molecules and their complexes. The timescales explored in MD simulations enable the study of sub-microsecond scale diffusion phenomena. A precise definition of the molecular simulation environment in which the diffusion properties of tRNA complexes are to be studied is an important prerequisite for MD simulation studies. Such a model cytoplasm system is not readily available for eukaryotes. Proteins and nucleic acids predominantly contribute to macromolecular crowding in the cytoplasm and characterization of diffusion properties in crowded systems requires simulation times of the order of a few microseconds. An MD simulation approach that can accommodate these two criteria is another prerequisite.

1.7 Aims of the study

The effect of macromolecular crowding on the diffusion of tRNAs and ternary complexes in eukaryotes is not well understood. A model yeast cytoplasmic environment that is essential to accurately characterize the diffusion properties using a molecular dynamics simulation approach is not available in literature. The origin of altered diffusion properties like slow and sub-diffusion in crowded conditions needs further investigation. In this study the diffusion properties of eukaryotic ternary complexes and tRNAs were characterized using yeast cytoplasm as the model system. Simulation of diffusional association (SDA)[48] was chosen as the appropriate MD simulation approach. As SDA (with soft-core repulsive, electrostatic desolvation, hydrophobic desolvation, and electrostatic energy terms) had never been tested before for a crowded environment with more than one type of protein solute, its robustness for simulating tens of solute species was carefully assessed. A yeast cytoplasm model was defined

following a rigorous statistical analysis of available proteomics datasets and single-cell observations. The diffusion properties affected by altered crowding, as a result of stimuli like osmotic stress, were studied by simulating a reduced model cytoplasm at high concentrations of proteins.

1.8 Structure of the thesis

This thesis is organized as follows.

Chapter 1: Introduction (current chapter)

Chapter 2 contains a detailed literature review of the studies on diffusion in the crowded environment (both experimental and computational). Various models of sub-diffusion and their properties are also discussed here.

Chapter 3 contains the details of the investigations into crowded protein solutions. The causal relations of slow and sub-diffusion in protein solutions are characterized in this chapter. The robustness of SDA in predicting these properties in solutions with multiple types of protein solutes is carefully established.

Chapter 4 details the process of development of minimal contents of the model yeast cytoplasm. Several yeast proteomic datasets were analysed carefully to determine the species and numbers of crowders to be incorporated in the model cytoplasm.

Chapter 5 contains the details of pre-processing of the yeast simulation cell contents and the results of the simulations.

Chapter 6 outlines the conclusions and future directions.

Chapter 2 Literature review

The research described in this thesis aimed to characterize the diffusion properties of tRNAs and ternary complexes under crowded conditions. The diffusion properties under crowded conditions are altered, resulting in slow and sub-diffusive behaviour. There are multiple explanations for both of these phenomena. More specifically, in the case of sub-diffusion, an understanding of the underlying stochastic process that mediate sub-diffusion is necessary to correctly interpret the data. The properties of different stochastic models cannot be used interchangeably. Therefore, in order to understand sub-diffusion in tRNAs and their complexes, an understanding of the features of such sub-diffusion is necessary.

Diffusion under crowded conditions has been characterized using *in vivo, in vitro,* and *in silico* approaches. Various experimental techniques have been employed to study diffusion and they vary in terms of time resolution, invasiveness of the procedure, and the ability to characterize both slow and sub-diffusion. On the other hand, diffusion can be characterized by molecular simulation approaches, which can also provide valuable insights into the causal relations behind the altered diffusion properties.

In this chapter, a brief recapitulation of the theory of diffusion and a review of models of sub-diffusion that can potentially be used to characterize the diffusion of tRNAs and ternary complexes is provided in the first section. The second section deals with experimental approaches used to characterize diffusion *in vitro* and *in vivo.* These approaches are reviewed with a focus on their suitability for characterizing the diffusion properties of tRNAs. In the final section computational studies of diffusion in monodisperse and polydisperse crowded solutions are reviewed, identifying gaps in knowledge.

2.1 Theory of diffusion and sub-diffusion

Albert Einstein, in his seminal work on Brownian motion[49], derived an equation for the diffusion coefficient of a particle in a system in dynamic equilibrium by balancing the flux due to diffusion and the flux due to a position-dependent force:

$$D = \frac{RT}{N_A} \cdot \frac{1}{6\pi\eta r}$$

Equation 2.1

where ‘D’ is the diffusion coefficient, ‘R’ is gas constant, ‘N_A’ is Avogadro’s constant, ‘η’ is coefficient of viscosity, ‘r’ is the radius of the particle, and ‘T’ is the temperature.

While describing the irregular thermal motion of particles, Einstein derived a relation between the diffusion coefficient and the mean square displacement of a particle. Equation 2.2 describes this relation for a particle diffusing in three-dimensional space. The term on the left-hand side of the equation is called the mean squared displacement (MSD)

$$< x^2(t) > = 6Dt$$

Equation 2.2

The assumptions made in doing so are that

(i) the movement of a particle is independent of that of the other particles in the system,

(ii) for a very small interval of time (compared with observation time) the consecutive displacements are independent of each other,

(iii) the particle follows a mean free path during the aforementioned small interval of time and the displacement during this time is small,

(iv) the distribution of displacements is symmetric i.e., the probabilities of displacements ‘x’ and ‘-x’ are the same.

Anomalous diffusion, where Equation 2.2 takes a power law form given by,

$$< x^2(t) > \sim t^{\alpha}$$

Equation 2.3

arises when one/more of the above assumptions is/are not satisfied. Depending on the value of α, the anomalous diffusion behaviour is referred to as sub-diffusive ($0<\alpha<1.0$) or super-diffusive ($\alpha>1.0$). In the context of biological systems, sub-diffusion can arise as a consequence of macromolecular crowding in cells. For example, sub-diffusion is observed in *in vitro* experiments of streptavidin[50], in the *in vivo* studies of membrane proteins[51], globular proteins in muscle cells[52], and fluorescent labelled polymers microinjected into cells.[53] Sub-diffusion can arise from different phenomena and multiple models (or their combinations) explain such anomalous behaviour. Some of the most popular models of sub-diffusion are discussed below.

2.1.1 Continuous time random walk (CTRW)

In CTRW the particle or the entity under consideration takes steps of size 'x' that follow a distribution d(x) and waits for a time 'τ' between steps. The waiting time 'τ' is not a constant and is chosen from a distribution g(τ). [54,55] When the characteristic waiting time, given by <τ>, and the second moment of d(x) are finite, the diffusion is normal and the process is ergodic.[56] In an ergodic process, the ensemble average of a property (for example, MSD) is equal to its time average. However, when g(τ) decreases according to a power law for large values of τ, it results in sub-diffusive behaviour with weak ergodicity breaking.[56] The waiting time(τ) varies between different visits of the same spatial point and this is called annealing.[57] The sub-diffusion of nanoparticles in the cytosol of mammalian cells is explained by CTRW. The sub-diffusion arises as a result of long-tailed waiting times arising from the non-specific interactions of the nanoparticles with the cytoplasmic components.[58]

2.1.2 Fractional Brownian motion (fBm) and fractional Langevin equation motion

fBm, developed by Mandelbrot and van Ness[59], has been recently used to explain the diffusion behaviour in confined crowded environment[60] and polysaccharide dextran crowded solutions.[61] The 'fractional' part of the fBm refers to the associated fractional Gaussian noise. In a discrete-time fBm, the total displacement of a particle at time 't' is given by, $x_t = x_{t-1} + \xi_t$. ξ_t corresponds to fractional Gaussian noise, with zero mean, which is correlated according to the function[62] given in Equation 2.4,

$$C(j) = < \xi_i \xi_{i+j} > = \frac{1}{2}\sigma^2\left(|j-1|^\beta - 2|j|^\beta + |j+1|^\beta\right)$$

Equation 2.4

where 'σ^2' is the variance of Gaussian noise distribution. fBm is a Gaussian process[63] without ergodicity breaking,[56] whereby the time averaged and ensemble averaged properties deviate from each other. The ensemble averaged MSD (EAMSD) is proportional to t^β where 'α' in Equation 2.3 is equal to 'β'. The nature of 'β' in Equation 2.4 determines if the system is persistent or anti-persistent. For β <1.0, the system is anti-persistent with C(j) being negative and when β >1.0 it is persistent with a positive C(j). Since the anomalous diffusion coefficient is equal to 'β', sub-diffusion in fBm is associated with an anti-persistent correlation function.

Recently, the behaviour of a particle executing fBm in the presence of reflective boundary at x=0, was characterized by Wada and Vojta.[62] In their simulations, the particle is bound by a reflective wall and for α deviating from 1.0, non-Gaussian behaviour is observed in the displacement distribution. The simulations conducted by Guggenberger et al.,[64] where a particle executing fBm in a finite interval, captured similar features. Figure 2.1 depicts the displacement distribution calculated for long time intervals, which shows deviations from Gaussian behaviour. These two examples show that although fBm is a Gaussian process, deviations from Gaussian are observed under certain conditions.

Fractional Langevin motion on the other hand constitutes fractional Gaussian noise coupled with friction kernel as given by[65] Equation 2.5

$$m\frac{d^2x(t)}{dt^2} = -\gamma^*\Gamma(\beta-1)\frac{d^{2-\beta}x(t)}{dt^{2-\beta}} + \eta^*\xi_{fGn}(t)$$

Equation 2.5

where γ^* is the friction coefficient and η^* is the noise amplitude. The equation considers that there are two forces acting on the particle: a frictional force and a force due to the Brownian motion of particles. The strength of the frictional force depends on the fractional derivative of the particle position with respect to time. EAMSD in the short time limit is given by[56]

$$EAMSD \sim \frac{k_B T t^2}{m}$$

Equation 2.6

where T is the temperature and 'm' is the mass of the particle. For long time limit however, the equation changes to[56]

$$EAMSD \sim 2k_B T(\Gamma(\beta-1)\gamma^*)^{-1}t^{2-\beta}$$

Equation 2.7

which resembles Equation 2.3, where the anomalous diffusion coefficient α=2-β, such that the system shows sub-diffusive behaviour for $1<\beta<2$, indicating that positive correlations in the fractional Gaussian noise give rise to sub-diffusion. This is in contrast to what is seen in fBm where an β>1.0 gives rise to super-diffusive behaviour. Some of the places where fractional Langevin equation is applied are, intramolecular motion in proteins, where the fluctuation of

the distance between an electron donor and acceptor pair is explained using FLE[66], and the anomalous diffusion of lipid molecules in a bilayer.[67]

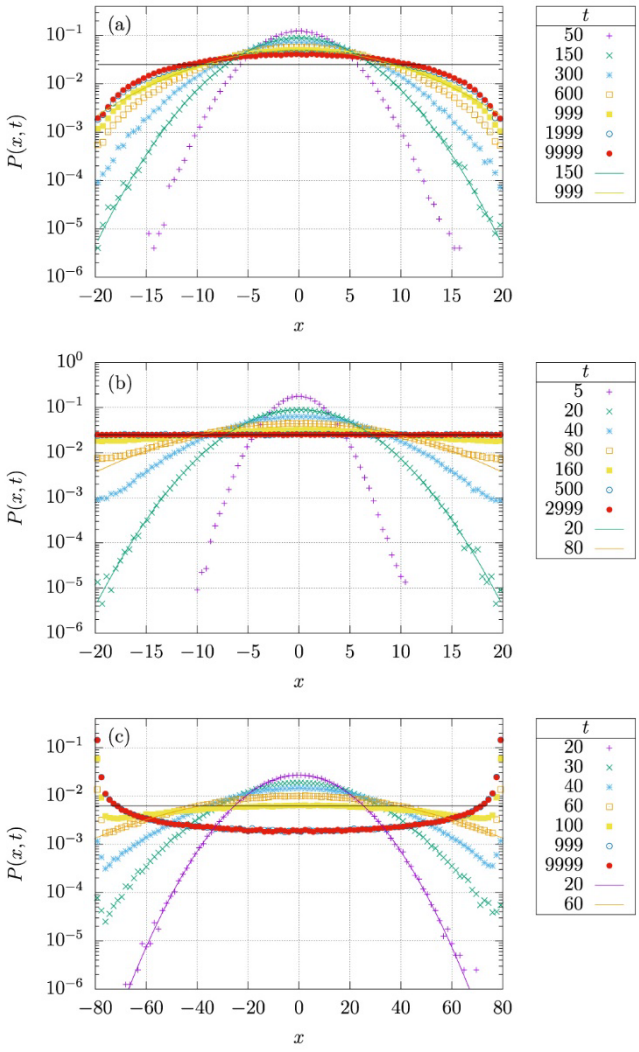

Figure 2.1 The displacement distributions of the simulations conducted using fBm in a finite interval under (a) sub-diffusive, (b)normal diffusive and (c) super-diffusive conditions. A clear deviation from Gaussian distribution is observed for long time simulations. The solid lines represent ideal Gaussian distributions at corresponding times. This figure was taken from the work of Guggenberger et al.[64] (doi: https://doi.org/10.1088/1367-2630/ab075f)

2.1.3 Other models or combination of models of anomalous diffusion

In contrast to fBm, 'scaled Brownian motion' and 'heterogeneous diffusion process', both show weak ergodicity breaking. In scaled Brownian motion the diffusion coefficient is modelled as a time varying quantity, whereas it is modelled as a space varying quantity in heterogeneous diffusion process.[56] Interestingly, there were experimental observations depicting normal Brownian motion with non-Gaussian distribution of displacements.[68] This

anomalous motion was later explained by invoking the concept of 'diffusing diffusivity', where the diffusion coefficient at every step is taken from a distribution indicating the random walk behaviour of diffusivity itself.[69]

In their experimental observations on the diffusion of insulin granules *in vivo*, Tabei et al.,[57] observed mixed features of fBm and CTRW. While the *p*-variation test,[70] which is used to delineate fBm from CTRW, favoured fBm, the tests for ergodicity showed CTRW like behaviour. This mixed behaviour is explained by a unified model where the individual steps are generated using a correlation function similar to that of fBm and the waiting times are generated using a distribution that takes power law form like that of CTRW.[57] The resulting model successfully explains the hybrid nature of the insulin granule diffusion observed in the experiments. A similar hybrid model combining diffusion on fractal like structures with CTRW was earlier proposed by Meroz et al.[71]

In summary, sub-diffusion arises as a consequence of different processes. An understanding of the underlying model is essential before ergodicity or Gaussianity of the displacement distributions is assumed. The experimental or simulation data can be analysed keeping in mind the key features of these models which can provide crucial insights into the causal relations of sub-diffusion.

2.2 Experimental approaches to characterizing diffusion under crowded conditions

Diffusion under crowded conditions is characterized experimentally using different types of techniques that differ in terms of the nature of data acquired, invasiveness of the approach and the timescales explored. Some of the popular techniques are fluorescence correlation spectroscopy (FCS), fluorescence recovery after photobleaching (FRAP), nuclear magnetic resonance (NMR), neutron backscattering, and single particle tracking (SPT) using fluorescent probes. The techniques and some of their applications are described briefly in the sub-sections below.

2.2.1 FRAP

FRAP is one of the foremost techniques used to monitor diffusion *in vivo.* In this approach, a cell expressing a fluorescent molecule (like green fluorescent protein) is exposed to a beam of

high intensity light which results in the loss of fluorescent activity in the exposed area. The fluorescence in the photobleached area is recovered by the means of the diffusion of fluorescent molecules from unirradiated parts of the cell. The rate at which this recovery occurs is a measure of diffusion. The characteristic diffusion time, τ_D is extracted from the fluorescence recovery rate and is the time at which the fluorescence recovered is half the maximum fluorescence attained after bleaching. In the simplest case of diffusion in two dimensions, 'τ_D' is related to the diffusion coefficient(D) by the following equation[72] (Equation 2.8):

$$\tau_D = \frac{\omega^2 \gamma}{4D}$$

Equation 2.8

'ω' in Equation 2.8 corresponds to disc radius for laser beam with circular disc profile or half-width at $1/e^2$ height for a beam with Gaussian intensity profile, and 'γ' is the correction factor that accounts for the difference between the user-defined and effective bleaching.[73] FRAP was later extended to study anomalous diffusion in crowded systems.[74] Even with the most recent advances like modified Line-FRAP approach[75], the lower limit in the timescales explored in FRAP is of the order of a few milliseconds. The time resolution can go up to one second depending on the microscope used.[76] FRAP can be easily used to characterize diffusion *in vitro*[75] or *in vivo* in membranes,[77,78] nucleus,[79,80] and cytosol.[81,82]

2.2.2 FCS

In FCS, the fluorescence intensity of tracer molecules is monitored as a function of time. The fluctuations in the fluorescence intensity, in the volume sampled by the beam, arise due to changes in the chemistry of the tracer and/or as a result of its translation motion. The autocorrelation of the intensity fluctuations, observed as a function of time, is calculated using Equation 2.9[83],

$$G(\tau) = \frac{\langle \delta I(t) \delta I(t+\tau) \rangle}{\langle I(t)^2 \rangle}$$

Equation 2.9

where I(t) corresponds to fluorescence intensity at time t and δI(t) is equal to I(t)-<I(t)>. The autocorrelation function calculated is fitted to a function that is dependent on 'τ_D' and using

Equation 2.10 for 'τ_D', the diffusion coefficient can be calculated by substituting beam waist (which is half-width at $1/e^2$ height for a beam with Gaussian intensity profile) for 'r_o'.[83]

$$\tau_D = \frac{{r_o}^2}{4D}$$

Equation 2.10

The function used to fit the autocorrelation data can be modified to accommodate anomalous diffusion. However, models corresponding to diffusion of multiple molecules also give rise to similar trends in the autocorrelation function, delineation of these two effects is therefore challenging.[83] In contrast to FRAP, the timescales explored in FCS are of the order of microseconds. The lower limit on the timescales is imposed by the deadtime of the detector and afterpulsing.[83] FCS has been used to study diffusion *in vitro* in crowded protein solutions[33] and *in vivo*.[84,85]

2.2.3 Pulsed field gradient NMR

Pulsed field gradient is used to study rotational and translational diffusion properties of molecules in crowded protein solutions.[86,87] A gradient pulse is used for spin spatial encoding, which creates a position dependent phase shift in the spins. The molecules are then allowed to diffuse for a period called 'diffusion time' and this is followed by a decoding pulse that is applied to nullify the phase shift of spins. In an ideal scenario where there is no diffusion the decoding pulse removes the phase shift perfectly, however due to diffusion of the molecules the phases are not reset, such difference in the expected and observed nature of spins acts as the signal for molecular motion. The 'diffusion time' allotted varies between a few milliseconds to seconds which is the timescale of diffusion explored in these experiments.[88] A more detailed description of these experiments is available in references [89,90].

2.2.4 Neutron backscattering

Neutron backscattering was employed to study the diffusion properties of bovine serum albumin (BSA)[91,92] and immunoglobulin tracers in the cell lysate solutions.[93] The set up requires deuterated crowding environment in which the tracer diffusion is studied. The relation between the intensity of the signal and energy (ω) is given by a scattering function. The scattering function may vary with respect to the magnitude of the scattering vector.[93] The scattering vector is the vector difference between the scattered wave vector and the incident

wave vector.[94] The scattering function obtained from the experiments can be expressed as a sum of three Lorentzian functions, corresponding to the internal motion of molecules (fast process), diffusion (slow process) and solvent contribution.[93] The width of the Lorentzian corresponding to diffusion provides information about the diffusion coefficient. The slope of the curve between the 'square of the scattering vector' and the 'width of the Lorentzian' is equal to the diffusion coefficient. [92,93] The deviations from the linearity of this curve are used to infer anomalous diffusion.[92] The timescales accessible vary from picoseconds to nanoseconds,[93] which is a lot smaller than that of the other techniques explored in this section.

2.2.5 Single particle tracking (SPT)

Single particle tracking is the most straightforward technique to measure diffusion coefficients and anomalous diffusion in cells. The technique dates back to Nordlund's observations, in 1914, leading to the collection of timeseries data on the motion of mercury droplets.[56] In recent times, single particle tracking typically involves fluorescence tagging of the target molecule and following its motion inside the living cells.[95] The technique has been applied to study the diffusion properties of tRNAs inside bacterial cells.[30] In this experiment, fluorescent labelled tRNA were transfected into the *E.coli* cells; tracking them (as shown in Figure 2.2) revealed a bimodal distribution of the diffusion coefficients. The authors attribute these diffusion coefficients to free tRNAs and tRNAs bound to ribosomes. The timescales explored in these simulations are of the order of milliseconds to seconds.

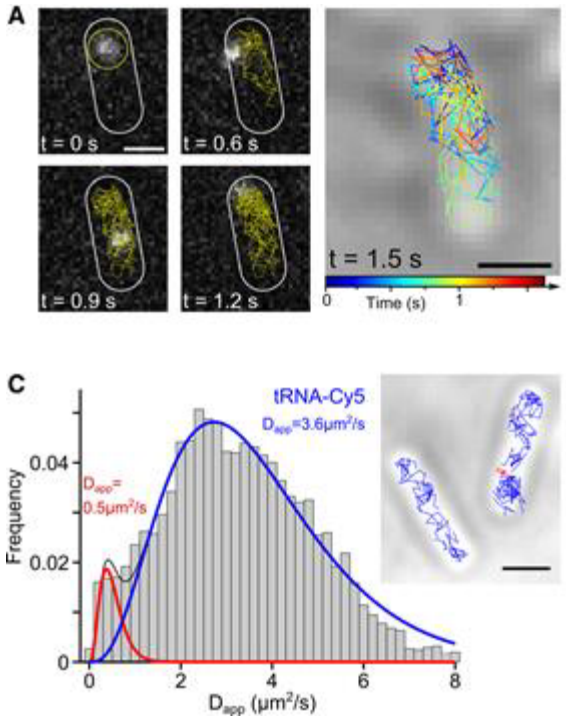

Figure 2.2 The trajectories of tRNAs in E.coli cells revealed from the tracking of fluorescently labelled tRNAs is shown in palette 'a' and palette 'c' indicates that the distribution of apparent diffusion coefficient arises from the diffusion properties of two species, the red line corresponding to ribosome bound slow species and blue line corresponds to fast-diffusing, presumably unbound tRNAs. This figure is taken from the work of Plochowietz et al.[30] (doi: https://doi.org/10.1093/nar/gkw787)

2.2.6 Discussion

The methods described above were used to describe the slow diffusion of proteins and polymers under a variety of crowded conditions (both *in vivo* and *in vitro*). However, the time scales explored in each of these methods varied. While FRAP, SPT, and pulsed field gradient NMR allowed exploration of milliseconds to second timescales, FCS and neutron backscattering facilitated exploration of phenomena at the microsecond and nanosecond time scales, respectively.

Konopka et al. investigated the diffusion properties of GFP in bacterial cells adapted to high osmolality conditions, and also cells that are subjected to plasmolysis (using NaCl) with the help of FRAP. Their experiments showed that the diffusion of GFP, under the highly crowded conditions of plasmolyzed cells, reduced by 70 times as the volume fraction of the macromolecules increased from 0.16 to 0.33.[96] FRAP has been extended to study diffusion in three dimensions where a volume of the sample is photobleached.[97] This technique was used

to study the anomalous diffusion of GFP in HeLa cells. While the diffusion of GFP in phosphate buffered saline and glycerol showed normal behaviour, anomalous diffusion was observed *in vivo*. This anomalous diffusion was attributed to the largely immobile obstacles present in the cells.[97] However, it is important to note that the FRAP signal obtained as a result of the reversible nature of photobleaching is similar to the one obtained for a fluorophore exhibiting anomalous diffusion.[97] Therefore care must be taken when inferring anomalous diffusion from FRAP experiments. The *in vivo* diffusion properties of the microinjected dextran molecules, studied using FCS, indicated that anomalous sub-diffusion in cells exists even after the disruption of cytoskeletal elements.[98] Variable-length scale FCS has been used to characterize the nature of sub-diffusion in agarose gels and dextran. While the sub-diffusion in agarose gels exhibited non-Gaussian displacements, Gaussian sub-diffusion was observed in dextran solutions.[99] However, sub-diffusion that may arise at sub-microsecond scales as shown in the simulations of Weiss et al.[98] cannot be captured by FRAP or FCS. The same holds true for SPT or pulsed field gradient based studies due to the lower limit on the accessible timescales.

Although most of the SPT approaches are confined to studying motion in two dimensions, recent advances in SPT like 3D-SMART[100] facilitated investigations of low concentration BSA solutions in three dimensions. Similar 3D SPT approaches like MINFLUX allowed monitoring the x, y, and z coordinates of a molecule *in vivo*.[101] Inferring sub-diffusion in SPT studies is straightforward, where the MSD calculated from the trajectory can be used to estimate the α-exponent. Autofluorescence of molecules (like elastin[102]) poses a challenge while studying the *in vivo* diffusion using FRAP, FCS or SPT (with fluorescently labelled probe). Neutron backscattering, on the other hand, does not have this drawback and allows characterization of sub-microsecond scale sub-diffusion. However, the upper limit for timescales monitored in this approach is of the order of nanoseconds.[93]

Pulsed-field gradient NMR was used to study the diffusion properties of CI2 in the presence of BSA, lysozyme, ovalbumin, ficoll, glycerol, and cell lysates. The diffusion of properties of CI2 in the presence of protein crowders resembled that of cell lysates, which was not the case for the polymer crowders.[87] The diffusion of CI2 was measured in the presence of 50-300 g/L concentrations of the protein crowders (lysozyme, BSA and ovalbumin). However, the mechanism behind the slow-diffusion observed in these experiments, in the time scale of milliseconds to seconds, is not well understood. There are multiple experimental techniques

that can be potentially employed to study the diffusion properties of tRNAs and ternary complexes. The timescales of these techniques vary as described above and a single experiment that can monitor the diffusion properties at the order of few nanoseconds as well as few microseconds is not readily available. This is necessary to investigate sub-microsecond scale anomalous diffusion associated with macromolecules in highly crowded environments.[33] Secondly, a novel approach needs to be developed to track both tRNAs and ternary complexes, as the ternary complexes lead to the formation of individual tRNAs and vice-versa, tracking them separately *in vivo* is a challenge. Although the experimental approaches provide valuable insights into the diffusion behaviour of these molecules, the causal relations, in terms of molecular interactions, of slow diffusion or sub-diffusion that might arise in these molecules cannot be easily understood with experimental approaches alone.

2.3 Computational approaches to characterizing diffusion under crowded conditions

Diffusion can be studied computationally using molecular dynamics (MD) simulations. Computational approaches provide insights into the phenomena occurring at timescales that are inaccessible to experiments. Moreover, computational studies also enable detailed investigations into the molecular processes that underlie modified diffusion characteristics. Computationally, diffusion in monodisperse solutions is studied extensively using protein or polymer crowders. Such studies, as detailed below, provide key insights into the causal relations of modified diffusion characteristics. Fewer studies looked into the effects of polydisperse media like cytosol. The first sub-section deals with computational studies of mono- or bi-disperse crowded media with a focus on understanding the role of molecular interactions and excluded volume effects. In the second sub-section, computational studies on cytosol like crowding are reviewed. This part of the review focuses on the characterization of the model cytosol system, details of the simulation approach, and the key findings of the studies. This is followed by a short discussion delineating the gaps in the current understanding while reflecting on the agreement/disagreement between different approaches.

2.3.1 Macromolecular crowding and diffusion

Crowding effect is studied computationally using different levels of description of the crowders. The entire protein or polymer crowder molecule can be represented as a single

sphere, and the properties of such a sphere remain the same across its surface. However, since the properties of a molecule, such as electrostatic potential, vary across its surface, this can be captured by embedding beads on the surface of the sphere. Proteins can also be coarse grained as a chain of beads, or a few atoms in a residue can be combined to form a bead whose characteristics are specific to the group of atoms. For example, Cheung et al. and Minh et al. used a spherical description of the crowders to study the properties of WW domain[103] and HIV protease[104] respectively. A similar crowder representation was used by Oh et al. to study the properties of polymers in the environment of crowders of different sizes.[105] The crowder particles are represented as spheres with beads in Elcock's studies on protein trapped in GroEL cage[106], Kurniawan et al.'s studies on the folding kinetics of β-hairpin,[107] and Trovato et al.'s studies on bacterial cytoplasm.[108] A combined approach with both coarse-grained and atomistic level of detail is employed by O'Brien et al. in their studies on the effect of nanoparticles on amyloidogenic proteins,[109] and in the simulations of trp-cage by Bille et al.[110] Some of the studies[32,33,111–114] conducted at atomistic level of detail focussed on the mechanisms underlying the diffusion behaviour observed in crowded systems.

Feig et al. have studied the diffusion properties of chymotrypsin inhibitor 2 (CI2) in the presence of bovine serum albumin (BSA) and lysozyme (LYS). The all-atom simulations with explicit solvent were set up with one molecule of CI2 and 8 molecules of the crowder at a concentration of 100 g/L.[32] Based on these simulations, run for 117 ns (CI2 in BSA) and 244 ns (CI2 in LYS), it is inferred that LYS interacts preferably with CI2 whereas BSA molecules mostly interact with each other. This knowledge is then used to explain the sub-diffusion of CI2 in the presence of BSA, and predominantly normal diffusion in the presence of LYS. The absence of extensive interactions between CI2 and BSA allows the manifestation of cage effects leading to anomalous diffusion.[32] In their all-atom simulations with villin, Nawrocki et al[115] explained the slow-diffusion observed in the crowded solutions by invoking the formation of transient clusters. The simulations were set up at different concentrations of villin and the maximum number of villins in the simulation cell is 64. The composition of clusters of different sizes in the simulations is calculated. The weighted-average of the diffusion coefficients of the representative clusters, calculated using HYDROPRO[28], was used to calculate the overall diffusion coefficients of villins. The agreement between the directly measured diffusion coefficients and those calculated from clusters showed high correspondence, reinforcing the role of transient clusters in slow diffusion. Similarly, in the simulations conducted by Bulow et al.,[113] dynamic cluster formation explains the slow diffusion in the dense protein solutions

of ubiquitin, third IgG-binding domain of protein G, villin head piece, and lysozyme. These monodisperse-all-atom simulations were performed at different concentrations of the solute with a maximum concentration of 200 g/L. It is important to note that in all of the above simulations, the diffusion properties investigated are in the time scales of few tens of nanoseconds.

On the other hand, atomistic detailed simulations conducted by Mereghetti et al[111,112,116] and Balbo et al.[33] explored the diffusion properties of crowded protein solutions in the microsecond timescales. The motion of macromolecules in the solution results in flows of solvent molecules. This movement of solvent molecules results in the movement of solutes in the vicinity. This coupling of the movements of macromolecules is termed 'hydrodynamic interactions' between the molecules. A mean-field approach was used to treat hydrodynamic interactions in these simulations. In the simulations by Mereghetti et al., the observed diffusion properties of lysozyme (at <60 g/L) agree well with that of a theoretical model that does not invoke cluster formation.[116] Based on their simulations with highly crowded myoglobin, hemoglobin A and hemoglobin S, it was concluded that shape effects, excluded volume effects, and hydrodynamic interactions play a major role in the slowing of diffusion. The simulations on haemoglobin A and IgG show that the absence of hydrodynamic interactions in the former resulted in deviations between the predicted and experimental diffusion coefficients, whereas the simulations of IgG did not show such behaviour.[111] This indicates that the importance of different types of interactions is dependent on the nature of the solute under consideration. In another simulation study with BSA and IgG by Balbo et al., transient anomalous diffusion was observed in the sub-microsecond time scales in highly concentrated solutions. The simulations conducted in the absence and presence of attractive forces showed that the sub-diffusion is less pronounced in the latter due to the formation of transient oligomers.[33]

The causal relations of this sub-diffusion occurring at the given timescales (in Feig et al. (2012) and Balbo et al. (2013)) have not been thoroughly investigated. Given that long-time diffusion coefficients are mostly affected by excluded volume effects, the role played by transient clusters or dynamic clusters needs to be further investigated in a manner specific to the protein type. The role played by the transient clusters in inducing or mitigating the anomalous sub-diffusive behaviour needs to be thoroughly investigated. The simulations described above deal with solutions of one or two protein types, an extensive review of the studies into the polydisperse, cytoplasm-like, crowded solutions is given below.

2.3.2 Cytosol like crowding and its effects

In this sub-section, the simulations conducted with model cytoplasm and the findings of these simulation studies are presented. This is followed by a discussion comparing the different types of simulation approaches, their evolution over time, and their limitations.

2.3.2.1 The Bicout and Field simulation approach

In 1995, Bicout and Field studied the effects of a polydisperse crowded medium using a computational approach for the first time.[117] Ribosomes, proteins, tRNAs and mRNAs were considered as the main constituents of their model cytoplasm, the composition of which is based on the details provided in Goodsell's paper[118] aptly titled, 'Inside a living cell'. A total of 12 ribosomes, 188 proteins, and 136 tRNAs, eventually chosen as the major ingredients, were represented as spherical particles. The densities of the atom types, hydrogen, carbon, nitrogen, oxygen and phosphorus were considered for calculating the interaction energies. All the proteins are treated as the same and therefore have the same atom densities. Such atom densities are also defined for tRNAs and ribosomes. In addition to this, the particles belonging to tRNAs, protein and ribosome molecules have different radii, diffusion coefficients, and mass density. The interaction energy between two particles is given as the sum of the Lennard-Jones and electrostatic term. Repulsive and dispersive interaction energy term between particles 'i' and 'j' is given by Lennard-Jones like potential. (Equation 2.11)

$$\varepsilon_{ij}^{LJ} = \int d\vec{r}_i \int d\vec{r}_j \left(\frac{A_{ij}^{LJ}}{r_{ij}^{12}} - 2.\frac{B_{ij}^{LJ}}{r_{ij}^{6}}\right)$$

Equation 2.11

Where A_{ij} takes the form,

$$A_{kl} = \sum_{\alpha \in k} \sum_{\beta \in l} \rho_k^{\alpha} \rho_l^{\beta} \sqrt{d_\alpha d_\beta} s_{\alpha\beta}^{12}$$

Equation 2.12

In Equation 2.12 the densities of different atom types in a particle (denoted by 'k' or 'l') are given by 'ρ', 'd' and 's' are the Lennard-Jones parameters. The DLVO (Derjaguin-Landau-Verwey-Overbeek) approach was used for calculating the electrostatic term.[119] The effect of counter-ions is accounted for implicitly in this approach and as a result, there is a very high

charge on the ribosomes. Multiple systems with varying charge on ribosome are set up to study the effects of such high charge on ribosome. The positions of the particles are evolved in time using Langevin approach, as the conditions required for the applicability of Ermak-McCammon equation are not met in the simulations, due to the steep potentials that require smaller time-steps. The Ermak-McCammon equation is derived by averaging Langevin equation in a time span of $\Delta t > mD/k_BT$, where m, D, k_B and T are mass, diffusion coefficient under dilute conditions, Boltzmann constant and temperature respectively. The simulations are run for 7.5 μs. The findings of their study primarily show a decrease in diffusion coefficients with an increase in the size of the particles and tRNAs are the fastest of the group. An interesting finding is that with an increase in the volume fraction (i.e. the volume ratio of the solute to the solution) the diffusion coefficients of tRNAs resembled the average of the diffusion coefficients of all the particles in the system, whereas the diffusion of proteins deviated from it. The findings of the variation of the diffusion coefficients with an increase in volume fraction should be taken with caution due to an associated change in the total charge of the system with an increase in the volume fraction.

2.3.2.2 The Ridgway et al. reaction-diffusion model

Ridgway et al. used a reaction-diffusion model approach to study the diffusion properties and reaction rates in crowded media.[25] The virtual cytoplasm in their approach considered 118 polypeptides (Figure 2.3). These polypeptides were described as monomers, homo- or hetero-complexes. The positions of the particles were evolved based on the diffusion coefficients estimated from the diffusion coefficient of green fluorescent protein (GFP). The particles were represented as spheres with a radius corresponding to its mass. Unlike the simulations of Bicout and Field[117], the forces between the particles were not calculated explicitly. This simplified approach allowed the study of properties on the scale of 10 μs. These simulations explained the diffusion-controlled association of barnase and barstar. Most importantly, their findings showed the presence of anomalous diffusion in crowded systems and a positive correlation between such anomalous behaviour and molecular mass.

2.3.2.3 The McGuffee and Elcock model of prokaryotic cytoplasm

Elcock and McGuffee's investigations into the crowding effects on diffusion constitute one of the most extensive studies in the field.[47] Fifty types of macromolecules in *E.coli,* contributing to nearly 85% of the total macromolecular mass of the cytoplasmic proteins, were chosen along

with GFP. Fourty five of these macromolecular species are proteins and the other five are RNAs or protein-RNA complexes(ribosomes). The molecules are represented as rigid structures with atomistic level of detail. The simulations are conducted with Brownian dynamics approach using Ermak-McCammon equation with a total simulation time of 15 μs. The electrostatic forces between the molecules are calculated using the effective charge model developed by Gabdoulline and Wade.[120] A 12-6 Lennard-Jones(LJ) energy term was used to describe the van der Waals, steric and hydrophobic interactions. The well depth parameter of LJ term was optimized to accurately predict the diffusion coefficients. These simulations explain the 10-fold decrease in the *in vivo* diffusion rate of GFP. One of the most interesting results is the sub-diffusion observed in proteins. The transient sub-diffusion reached a maximum between 10-100 ns with normal diffusion prevailing in the timescales above and below that range. Interestingly, higher values of α-exponent are observed in the presence of repulsive ($1/r^{12}$) only interactions hinting at the role of attractive interactions in inducing sub-diffusion.

2.3.2.4 The Wang and Cheung study of a coarse-grained cytoplasm

In this study by Wang and Cheung, the cytoplasm described in McGuffe and Elcock's work[47] was taken and coarse-grained further depending on the shape and size of the molecule. The underlying assumption in this approach is that the level of detail necessary to describe a crowder depends on the size of the tracer (or molecule of interest). If the tracer is larger than the molecule of interest no further coarse graining of the molecule is necessary else the molecule is coarse-grained accordingly. The cytoplasm environment defined in the previous sub-section was coarse-grained and replica exchange simulations were carried out using the molecular dynamics program AMBER10, employing Langevin equations of motion to characterize the thermodynamic properties of apoazurin. The structure of the coarse-grained crowders was maintained by defining bond, bond angle and dihedral angle energies between the beads of the crowders.

2.3.2.5 The Hasnain et al. model of cytoplasm

In this model[121] of *E.coli* cytoplasm, 159 protein species were represented using single or multiple beads depending on the size of the protein (Figure 2.3). Stretching and bending energies, modelled using harmonic potentials, constitute intramolecular interactions whereas the intermolecular interactions were modelled using a harmonic potential, the purpose of which was to avoid protein overlaps. This harmonic potential constituted the only type of inter-

molecular interaction apart from hydrodynamic interactions. Langevin dynamics were used to simulate this system with a cubic simulation cell of 406 Å edge length. The simulations predict the *in vivo* diffusion coefficient of GFP with a high degree of accuracy. Anomalous diffusion is observed in the timescales of 10 ns-1μs. The presence of fBm was inferred from the displacement auto-correlation function. However, other features of fBm like Gaussianity of the displacements and ergodicity were not thoroughly investigated. Secondly, the possible role of other sub-diffusive models or combinations of sub-diffusive models as described in the earlier sections was not explored. Despite these caveats, the simulations provided insights into the possible presence of fBm like sub-diffusive behaviour in polydisperse crowded media.

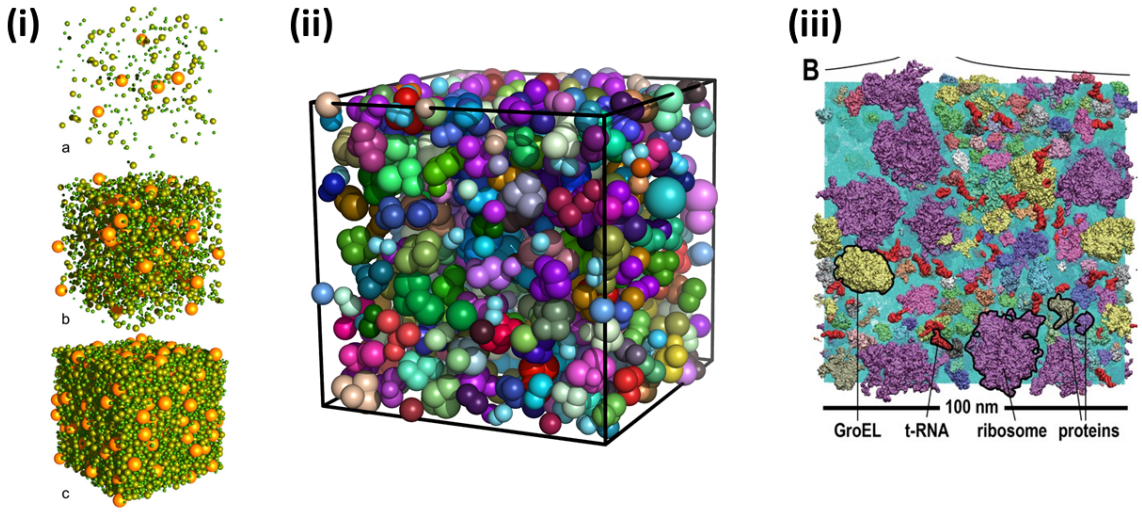

Figure 2.3 Evolution of the detail at which model cytoplasm is represented over the course of years. (i) Cytoplasm model of Ridgway et al.[25] at three different volume fractions represented by a, b and c. (Appendix: figure permissions 1). (ii) Hasnain et al.'s cytoplasm model[121] with macromolecules represented as spheres or cluster of spheres. (doi: https://doi.org/10.1371/journal.pone.0106466) (iii) Palette B cropped out of figure 1 in Yu et al.'s paper.[114] The figure shows macromolecules represented at the atomistic level of description with explicit solvent. (doi: https://doi.org/10.7554/eLife.19274.001)

2.3.2.6 The Trovato and Tozzini SpoB model

In Trovato and Tozzini's approach[108], 12 crowders were used to represent a cytosol-like model. The crowders were represented as a sphere of beads (SpoB). The beads were either hydrophobic or polar, contributing to soft and non-specific interactions, or purely repulsive interactions, respectively. The beads interacted according to a single-well potential, the parameters of which were obtained using a top-down approach from the diffusion properties of GFP. The interactions between the spherical particles were given by the sum of the pairwise

interactions between the beads on the particles. This sum was calculated for different orientations of the particles and the average of the sum was the interaction energy between the particles. The diffusion coefficients were calculated in the timescale of 5-20 μs. The simulations were carried out with and without attractive interactions. Interestingly, in the presence of attractive forces the system showed more pronounced sub-diffusion, consistent with the findings of McGuffee et al.;[47] however, the timescales of sub-diffusion varied between these models. The intensity of sub-diffusion varied depending on the size of the crowder, as shown in Figure 2.4. There is a clear correlation between the increased sub-diffusion and non-ergodicity, as inferred from Figure 2.4, hinting at non-ergodic origins of sub-diffusion. Although the ergodicity breaking (EB) parameter (Equation 2.13) of the nucleoid indicated the absence of ergodicity breaking in their simulations, it is important to note that the average TAMSD, calculated by averaging TAMSD over the particles of same type, used in the calculation of the EB parameter was only averaged across three particles.

$$EB\ parameter = 1 - \frac{TAMSD}{average\ TAMSD}$$

Equation 2.13

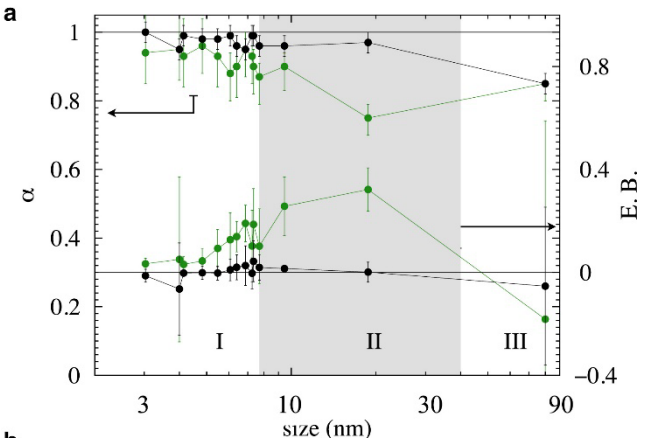

Figure 2.4 Palette 'a' of figure 6 in Trovato et al.'s paper.[108] The figure shows the variation of α-exponent and ergodicity breaking parameter(EB) with respect to the size of the crowder. The system is ergodic when EB is 0 and is non-ergodic when EB=1. A clear correlation between increase in non-ergodicity and sub-diffusion can be seen. The green colored lines correspond to simulations with attractive and repulsive interactions, whereas the black colored lines show the data from simulations with repulsive-only interactions. The three regions in the plot labelled as region I, II, and III correspond to particles of different sizes. Region II corresponds to particles of the size of ribosomes, region I and III correspond to particles that are smaller (most of the protein molecules) and larger than ribosomes (nucleiod) respectively. (Appendix: figure permissions 2)

2.3.2.7 The Ando and Skolnick cytoplasm model

The representative cytoplasm of *E.coli* developed by Ando and Skolnick[26] consists of 15 species of proteins and tRNA. The molecules are represented in two different ways: (i) each macromolecule is represented by a single sphere, (ii) C-alpha atoms and P, C4′, N1, and N9 in the tRNAs are treated as a single bead each. The simulations were conducted using Stokesian dynamics which facilitated easy implementation of hydrodynamic interactions. The rate of change of velocity of a particle in Stokesian dynamics simulations is given by the sum of the forces arising from hydrodynamic interactions, intermolecular interactions, and Brownian motion (due to random movement of particles). The intermolecular interaction energy is described by a Lennard-Jones potential. The simulations revealed that at very high concentrations of the crowder, representation of a macromolecule as a single sphere is sufficient to capture the diffusion properties. The reduction in the diffusion coefficient as a consequence of crowding was explained by hydrodynamic interactions alone. However, the role of non-specific interactions was not conclusively described in these simulations and requires further investigations.

2.3.2.8 The Feig et al. cytoplasm model and simulations

This is the most recent cytoplasm model developed by Feig and co-workers[122]. The model represents the cytoplasm of *Mycoplasma genitalium* and, so far, is the most extensive and detailed model. The model consists of 992 metabolic proteins, 40 ribosomes, 76 translation factors, 42 aminoacyl synthetases, 275 tRNAs, 48 RNA polymerases and other proteins represented in atomistic detail (Figure 2.3). The model also includes metabolities, ions, and a total of 26 million water molecules placed in a cubic simulation cell of length 100 nm, corresponding to $1/10^{th}$ of the size of the whole cell. The structures of the proteins in the model were obtained using homology modelling with the protein structure prediction program MODELLER.[123,124] In a follow up paper published in 2016, Yu et al[114] have subjected the above system to molecular dynamics simulations. The simulations were performed using GENESIS[125] using CHARMM c36[126] forcefield parameters for proteins and RNA, and CGenFF[127] based parameters for metabolites. Using the entire system described above, the simulations were run for 20 ns. Two further smaller systems ($1/8^{th}$ of the original system) were run for 60 and 140 ns each. Simulations were also set up with spherical representation of the particles with only repulsive forces using Brownian dynamics and Stokesian dynamics approaches. The diffusion coefficients measured in the Stokesian dynamics simulations with

hydrodynamic interactions showed good agreement with those of the atomistic simulations. It is interesting to see that the diffusion coefficients in crowded environment showed 1/R dependency, mimicking behaviour under dilute conditions. ('R' is the Stokes radius)

2.3.3 Discussion

In all of the models described above, microsecond scale timescales were explored, with the exception of Yu et al.'s all-atom simulations of *Mycoplasma* cytoplasm. The sub-diffusion observed in Trovato et al.'s and McGuffee et al.'s simulations can potentially be explained by CTRW-like behaviour. This can be understood by interpreting the non-specific interactions as the cause of long-tailed waiting times associated with CTRW. Since the sub-diffusion was more pronounced in the absence of attractive interactions[108,128] and is weakly non-ergodic,[108] this hypothesis is reasonable. In contrast to this, the sub-diffusion observed in Hasnain et al.'s model showed features of fBm, although a rigorous investigation into the sub-diffusive model was not carried out in this study. This lack of consistency needs to be investigated further, although the timescales at which sub-diffusion was observed in each of these simulations varied greatly. While hydrodynamic interactions (HI) were treated explicitly, due to the use of explicit solvent in Yu et al.'s simulations[114], a mean-field approach was employed by Hasnain et al.,[121] and Stokesian dynamics with the implementation of hydrodynamic interactions was used by Ando and Skolnick.[26] While hydrodynamic radii were used in the description of spheres in Bicout and Field's model, hydrodynamic interactions were not treated either explicitly or implicitly, and the same applies to other models. The trends observed in the diffusion coefficient vs Stokes radius curves in the all-atom model[114] mirror those in the simulations conducted by Ando and Skolnick (using spherical particles) with van der Waals forces (without HI). However, in terms of the absolute values of the diffusion coefficients in atomistic simulations, Stokesian dynamics with HI compare better, indicating the importance of both non-specific interactions and HI.[114] Recently, Grimaldo et al.[93] conducted simulations using Stokesian dynamics at two different levels of polydispersity index (PDI) corresponding to the model cytoplasm of McGuffee et al[47]. and Ando et al.[26], the former being high in PDI. Their simulations indicate that if the size of the tracer is close to the average size of the particles in a polydisperse medium, then the diffusion of the tracer is similar to that in the monodisperse media.

In all of the simulations described above, the cytosol corresponds to prokaryotic systems. The evolution of the resolution of model cytoplasm from macromolecules represented by spheres[117] to a cluster of spheres[121] and eventually to atomistic level resolution[114,122] is interesting to note (Figure 2.3). One can also notice the increasing complexity of the model cytoplasmic composition over the years. Despite a significant improvement in the models of bacterial cytoplasm in the past 20 years, there are no equivalent studies looking into the properties of the eukaryotic cytosol. It is also important to note that more comprehensive proteomics dataset for eukaryotes like yeast were only reported recently.[129] In addition, none of the studies mentioned above focussed on the sub- and slow-diffusive properties of tRNA ternary complexes (containing tRNAs, elongation factors and GTP). Although Feig et al.'s model accounted for elongation factors and McGuffee et al.'s included different types of tRNAs, the time scales explored in the former (few 10s of nanoseconds) and the absence of explicit treatment of ternary complexes in both, leave gaps in the understanding of translation mechanisms even in bacterial systems. A model eukaryotic cytoplasm, similar to Feig's cytoplasm model,[122] rigorously constructed using proteomics data, with which MD simulations can be readily performed, is not available. Given the importance of hydrodynamic interactions,[26] the importance of accounting for the polydisperse nature of the cytoplasm,[93] the role played by non-specific interactions in effecting sub-diffusion in a model cytosol,[47,108] an optimized molecular dynamics approach to study the diffusion in the eukaryotic cytosol in the microsecond time scale accounting for these effects has, to our knowledge, not been reported thus far.

Chapter 3 Characterization of slow and sub-diffusive behaviour in crowded protein solutions and discerning the underlying causal relations

3.1 Introduction

A computational approach that allows the simulation of multiple copies of RNAs, including tRNAs with modified bases, and different protein species simultaneously is necessary to accurately characterize diffusion in a model cytoplasmic environment. Simulation of diffusional association (SDA) facilitates such a study, at microsecond timescales, under implicit solvent conditions using Brownian dynamics. In this approach, the molecules are represented as rigid structures at an atomistic level of detail. This method has earlier been applied to study the diffusion properties of the monodisperse crowded solutions of bovine serum albumin (BSA)[33], γ-globulin[33], T4 and hen-egg-white-lysozymes[116], bovine pancreatic trypsin inhibitor (BPTI)[116], myoglobin[112] and, haemoglobin A[112] and haemoglobin S[112]. This approach has also been applied to understand the molecular mechanisms underlying the formation of nanoparticle-protein aggregates.[130] However, the method has not been tested for multiple species of proteins at a range of concentrations. Here, the diffusion properties of chymotrypsin inhibitor 2 (CI2) in the presence of bovine serum albumin (BSA) or lysozyme (LYS) were investigated using SDA and compared with predictions from all-atom MD simulations as well as experiments.

Wang et al. characterized the diffusion of CI2 under dense crowder conditions that varied from 50 to 300g/L, using NMR.[87] Additionally, atomistic simulations showed that chymotrypsin inhibitor 2 (CI2) exhibits significant sub-diffusive behaviour in the presence of bovine serum albumin (BSA) as the crowder protein, whereas no appreciable sub-diffusive behaviour was observed in the presence of lysozyme.[32] The absence of sub-diffusive behaviour in the lysozyme crowded environment was attributed to stronger interactions between lysozyme and CI2. However, it is important to note that the spatiotemporal scale explored in these simulations is relatively small, whereby a single molecule of CI2 was simulated in the presence of eight protein crowder molecules for 117-244 ns at a crowder concentration of 100 g/L. Nawrocki et al. later explained the sub-diffusive behaviour of CI2 by invoking cage effects.[115] 'Cage effects' arise when molecules are trapped in a transient cage formed by the surrounding molecules and

exhibit back-and-forth (rattling) motion, and the intensity of such an effect can be quantified.[131,132] However, to the best of our knowledge, no attempt has yet been made to quantitatively associate cage effects and sub-diffusive behaviour in crowded protein solutions. Since SDA facilitates microsecond-scale simulations at all the experimentally studied concentrations, sub-diffusion associated with these solutions and its origins were investigated. The consistency of the predictions of SDA-based simulations with the molecular interactions characterized in the atomistic simulations were carefully investigated. The findings of this analysis provided insights into the robustness of this approach for cytoplasm-scale simulations.

3.2 Approach and methods

The experimentally determined 3D structures of BSA (PDB: 3V03), CI2 (PDB: 2CI2) and LYS (PDB: 1AKI) were obtained from the Protein Data Bank (PDB). The Simulation of Diffusional Association (SDA, version 7.2.2) program was used to conduct Brownian dynamics simulations[48]. Pre-processing of the proteins, as described below, was done with webSDA[133]. The protonation states of amino acids in all proteins were assigned assuming a pH of 5.4 in order to emulate experimental conditions. Atomic charges and radii were taken from the AMBER force field 99.[134] Electrostatic grids of 1.0 Å resolution were calculated assuming an ionic strength of 200 mM (to also reproduce experimental conditions), with an ion radius of 1.5 Å, a protein dielectric constant of 4.0, a solvent dielectric constant of 78.0, and a temperature of 300 K, using the linearized Poisson-Boltzmann equation approach.[135] The electrostatic grids of LYS and CI2 were 129 x 129 x 129 $Å^3$ in size and the grid size of BSA was 193 x 129 x 161 $Å^3$, reflecting the differences in size and shape of these proteins. Effective charges were calculated using webSDA. Electrostatic desolvation, hydrophobic desolvation and Lennard-Jones energy grids were calculated at a resolution of 1.0 Å. The grid sizes of the electrostatic desolvation and Lennard-Jones (repulsive) energies of BSA, LYS and CI2 were 133 x 92 x 109 $Å^3$, 45 x 55 x 67 $Å^3$, and 43 x 44 x 45 $Å^3$, respectively. The size of the hydrophobic desolvation energy grids of BSA, LYS, and CI2 were 104 x 76 x 87 $Å^3$, 45 x 52 x 60 $Å^3$, and 44 x 44 x 45 $Å^3$, respectively. The energy grid files obtained were then used to set up simulations with protein crowder concentrations of 50, 100, 200 and 300 g/L, with CI2 as the tracer. Initial configurations were generated using the *genbox* tool in SDA by placing the proteins randomly in a cubic box of 350 Å length. To account for the potential influence of the

initial configuration of the proteins in each system, three systems with different initial configurations were set up for every concentration.

The simulations were performed using SDAMM (program used for simulations with multiple molecules) in SDA with a time step of 0.5 ps at the default SDA temperature of 300K. Each of the simulations was run with the softcore repulsive term only for one microsecond in order to remove any protein overlaps. The simulations were then run for one microsecond with the full energy term for equilibration purposes, followed by 5 microseconds of production runs. The self-diffusion coefficients of BSA and LYS were monitored to evaluate convergence, which was reached before one microsecond. Since both these crowders are larger than CI2, the convergence of the diffusion of BSA and LYS was expected to be slower and hence was used in this evaluation. Diffusion coefficients were calculated from the plots of time-averaged MSD (TAMSD) (obtained by averaging over all possible time origins) vs time (lag time). The simulations with the soft-core repulsive term (decaying at a rate of $1/r^6$) only were performed using the same approach as above except that both the equilibration and production runs did not include attractive interactions (the scaling factor of electrostatic, electrostatic desolvation, and hydrophobic desolvation terms is set to zero) in the energy term. The trajectories were unfolded assuming that any given particle does not move more than half the simulation cell length between time frames considered.[136]

3.2.1 Calculation of the α-exponent

The value of the α-exponent was calculated from the *log(TAMSD/τ) vs log(τ)* curve using an approach similar to that of Balbo et al.[33] Since the α-exponent is a time-varying quantity in our simulations, the straight-line region of the plot is chosen by fitting the parts of the curve to a linear fit in such a way that the R^2 value is maintained above a cut-off of 0.95. The regions at long timescales usually showed high levels of noise, which affected the quality of the fit. This is due to the use of TAMSD in our calculations, such that the MSD calculation is affected at large lag time values due to poor statistics. Therefore, long time scale regions with poor statistics were omitted from the calculations of the α-exponent. The average of the α-exponents calculated using the data from the three different initial configurations is reported below, and the error bars in the plots correspond to the standard deviation (STD), and p-values are calculated using two-tailed t-tests assuming unequal variance. It is important to note that, since log plots are used, the data at long timescales is crowded in a small region of the graph and, as

a result, while one can reliably calculate diffusion coefficients up to the order of a microsecond (in the TAMSD vs τ plots), it is not feasible to do a similar calculation of the α-exponent at long timescales with a stringent cut-off. However, since the α-exponent converges back to normal diffusion values within the range of timescales explored, this does not have any impact on our conclusions.

3.2.2 Quantification of cage effects

Cage effects were quantified using Doliwa and Heuer's approach.[132] Here, the displacement vector of a particle is given by $r_{mn}(\tau) = r(n\tau) - r(m\tau)$, where $r(n\tau)$ and $r(m\tau)$ are the position vectors at corresponding time points. The first and second displacement vectors are therefore termed r_{01} and r_{12}, respectively. The component of r_{12} along r_{01} is termed x_{12}. The component of r_{12} along an arbitrary vector perpendicular to r_{01} is termed y_{12}. According to this approach, it is expected that x_{12} be negative and decrease linearly with an increase in the magnitude of r_{01} in the presence of caging effects. This anti-correlation is due to the rattling motion of the particles. The vector y_{12} acts as control since it is the component along an arbitrary vector, so it would be expected that in the absence of caging, y_{12} and x_{12} exhibit similar behaviour upon the increase in the magnitude of r_{01}.[131,132] (Figure 3.1) The $|r_{01}|$ vs x_{12} (or y_{12}) plot is obtained by calculating the values of $|r_{01}|$ and x_{12} (or y_{12}) across all possible time origins along the length of the trajectory for a given protein, and combining the data for all the protein molecules (of a given species) in the simulation. The values of ($|r_{01}|$) are binned with a width of 0.05 Å and the corresponding x_{12} values are averages. The plots are presented and discussed in Figure 3.2. The approach described here is similar to that of the previous workers. The function used by Weiss[61] to infer anti-correlation is given by

$$C_\tau(t) = \left\langle \frac{v_\tau(T)}{|v_\tau(T)|} . \frac{v_\tau(T+t)}{|v_\tau(T+t)|} \right\rangle_T$$

Equation 3.1

where $v_\tau(t)=r(t+\tau)-r(t)$, 'r' being the position vector. When t= τ, this function is equivalent to the dot product of r_{01} and r_{12}, which is proportional to x_{12}. In the presence of anti-correlation, $C_\tau(t) < 0$ when t~ τ, which implies that x_{12} is negative, which is consistent with the above approach.

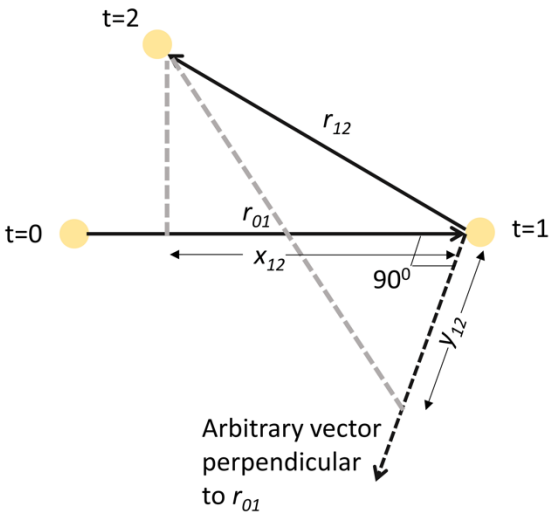

Figure 3.1 Quantification of cage effects. The particle is represented in yellow. Arbitrary vector perpendicular to r_{01} is represented as a dotted vector.

3.3 Results

3.3.1 Diffusion coefficients and sub-diffusive behaviour

Time-averaged translational diffusion coefficients of CI2, LYS and BSA were calculated from the curve of time-averaged mean squared displacement (TAMSD) vs time, averaging over all time origins and the molecular species of interest. The experimental diffusion coefficient of CI2 at concentrations of 50, 100, 200, and 300 g/L of LYS and BSA had been determined previously[87]. The long-time diffusion coefficients were calculated in the 0-1000 ns time range and the predicted diffusion coefficients of CI2 were compared with experimental values. Figure 3.2A shows that the predicted diffusion coefficients are of the same order of magnitude as experimental ones. However, the difference in the predicted and experimental diffusion coefficients increased at higher concentrations. This could potentially be due to the lack of flexibility in the protein structures, which could contribute to a reduction in the tendency to form clusters. A more detailed description of the role played by such clusters is provided further below. Figure 3.2B shows that the predicted diffusion coefficients of BSA and LYS decrease in magnitude with an increase in the concentration of the crowder, as expected.

The sub-diffusive behaviour of the proteins was characterised. In solutions with a crowder concentration of 50 g/L, the α-exponent value of CI2 hovered above 0.95 in the presence of both crowding proteins (Figure 3.3A and Figure 3.3C). The same behaviour was observed for the self-diffusion of the crowders, as shown in Figure 3.3E and Figure 3.3G. The α-exponent did not exhibit pronounced variation with respect to lag time in each of the systems. The increase in the concentration of the crowder led to sub-diffusion. At a crowder concentration of 300 g/L, the value of the α-exponent decreased to 0.83 (STD = 0.002) in the range 10.4-38.8 ns for CI2 in BSA, 0.87 (STD = 0.002) in the range 2.0-9.8 ns for CI2 in LYS, 0.74 (STD = 0.005) in the range 8.0-39.8 ns for BSA, and 0.80 (STD = 0.001) in the range 2.0-10.0 ns for LYS, in all cases indicating the presence of sub-diffusive behaviour. However, the observed sub-diffusion was transient and normal diffusion was gradually reached after a few hundreds of nanoseconds. In all of these cases a clear trend can be discerned, whereby diffusion is normal at short time scales, sub-diffusive in the sub-microsecond time scale, and back to normal at longer time scales. This behaviour is observed in all the three proteins, which are of different sizes and have a different total charge. An intermediate behaviour was observed in crowder concentrations of 100 and 200 g/L (Figure 3.4). Such transient sub-diffusive behaviour has been predicted for γ-globulin and BSA self-crowded solutions.[33]

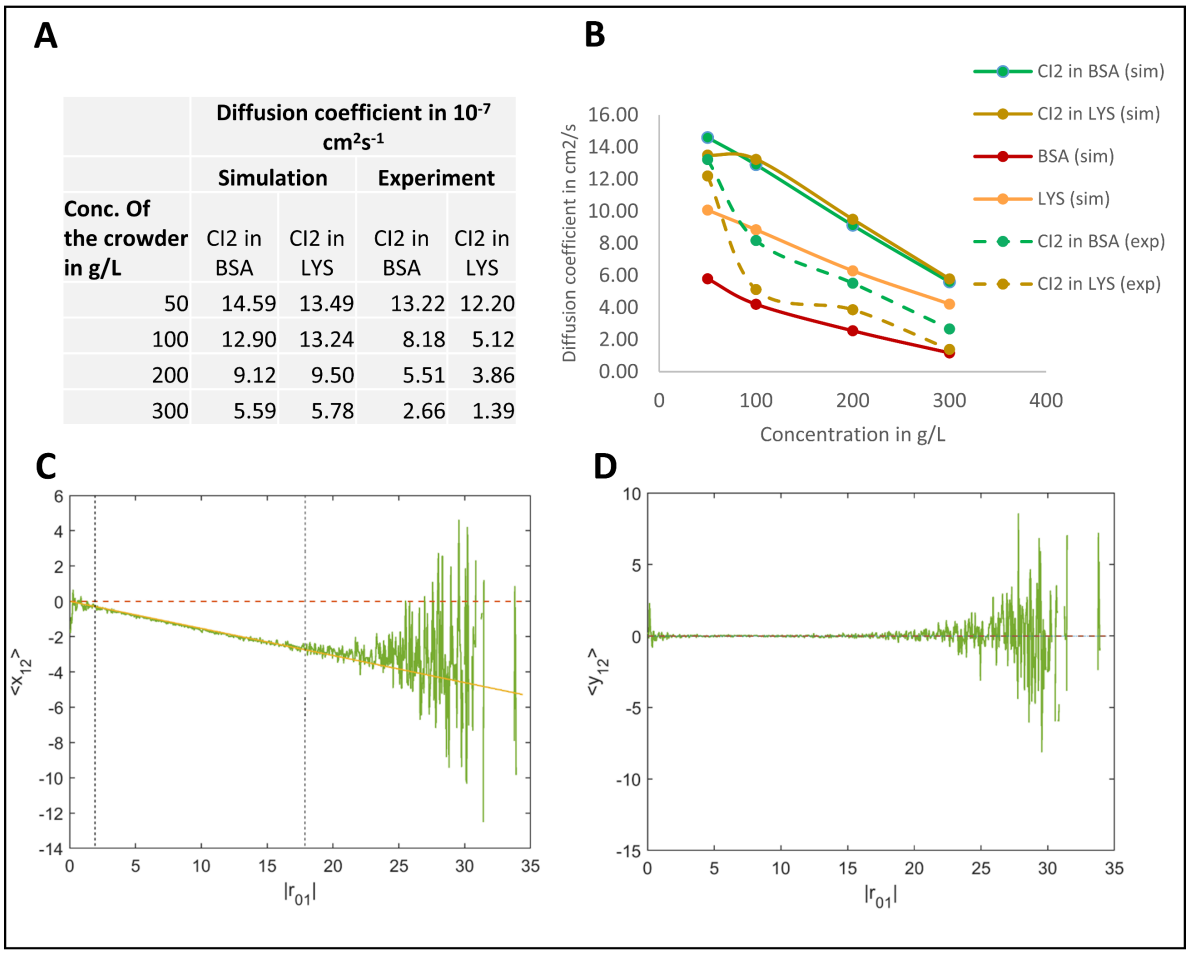

Figure 3.2 Predicted diffusion properties in crowded protein solutions. (A) Comparison of experimental and predicted CI2 diffusion coefficients. The predicted values are within the same order of magnitude of experiment, revealing good agreement. (B) The predicted and observed diffusion coefficients of the CI2 tracer in the presence of the protein crowders BSA and lysozyme, and of the protein crowders themselves are plotted as a function of crowder concentration. As expected, the increase in crowder concentration results in a downward trend of the diffusion coefficient of CI2. (C) Average x12 as a function of |r01| (green), whilst the dashed red line corresponds to the reference x = 0 curve, and the dotted vertical line separates the regions of low and high noise. The yellow line corresponds to the linear fit for the less noisy region, whose slope is used in the calculation of α-exponent. The slope is negative, indicating the presence of caging effects. (D) Average y12 as as a function of |r01| (green), whilst the dashed dotted line corresponds to the reference y = 0 curve, and the blue dotted line (which is very close to the y = 0 curve) corresponds to the linear fit of the less noisy region. |r01|, x12 and y12 are all provided in Å. Plots C and D correspond to data at a BSA concentration of 300 g/L at 5ns.

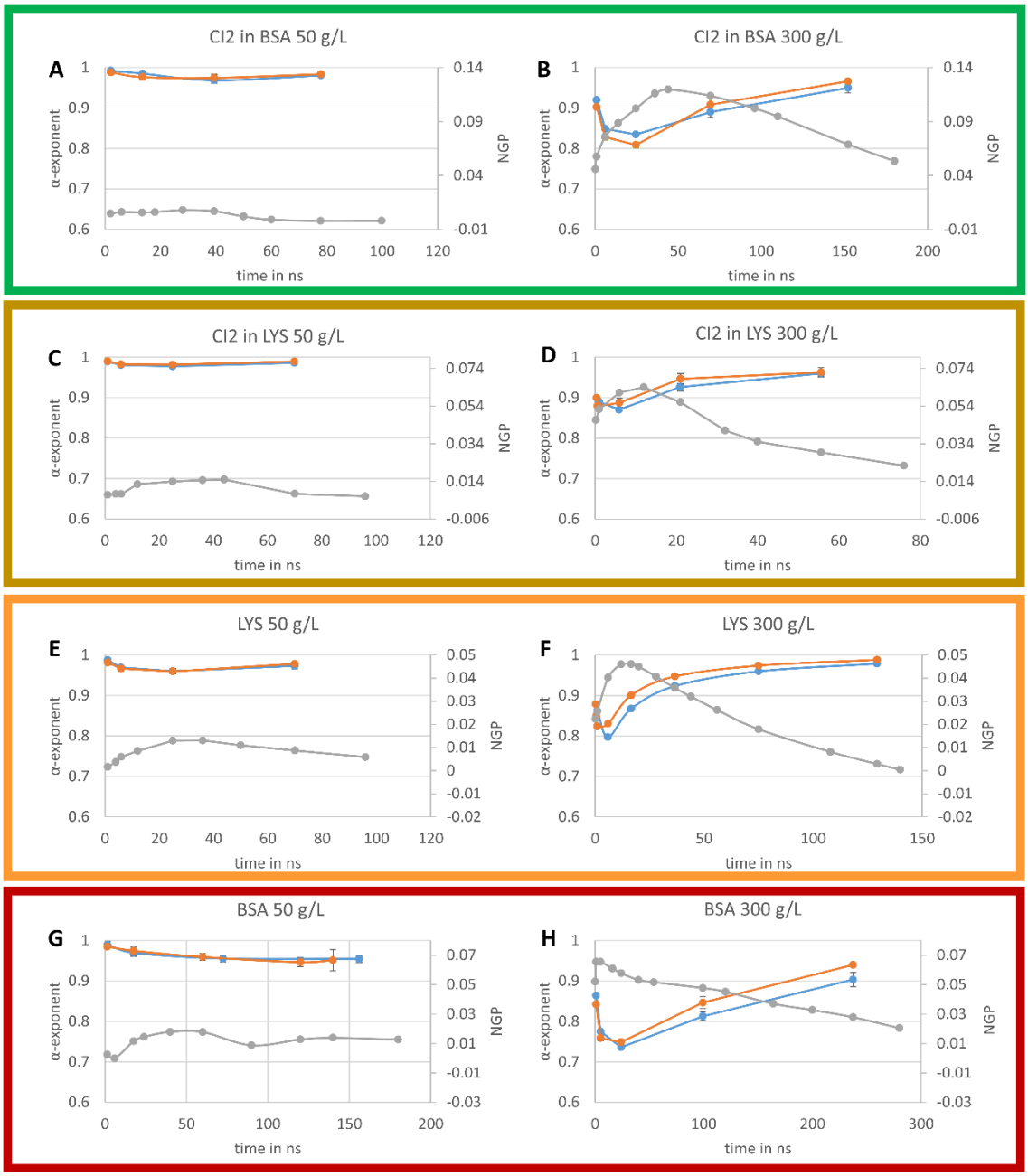

Figure 3.3. Sub-diffusive and non-Gaussianity properties of the crowders and tracer (50 and 300g/L concentration of crowder). The blue, orange and grey lines in all the curves represent the α-exponent calculated from the log(TAMSD/τ) vs log(τ) curves, α-exponent calculated from cage effects, and non-Gaussianity parameter (NGP) measured at different lag times respectively. All the curves on the left side of the figure represent the data for low concentration of the crowder at 50g/L and the ones on the right side represent data for high crowder concentration. The data for CI2 in BSA is in the first row highlighted in green, followed by data for CI2 in LYS in next row highlighted in yellow, followed by data for LYS and BSA highlighted in orange and red respectively. Error bars represent the standard deviation of the value of the α-exponent between simulations started with different configurations. The time ranges in the individual graphs are different from each other due to the variation in the emergence of noise in the log(TAMSD/τ) vs log(τ) curves.

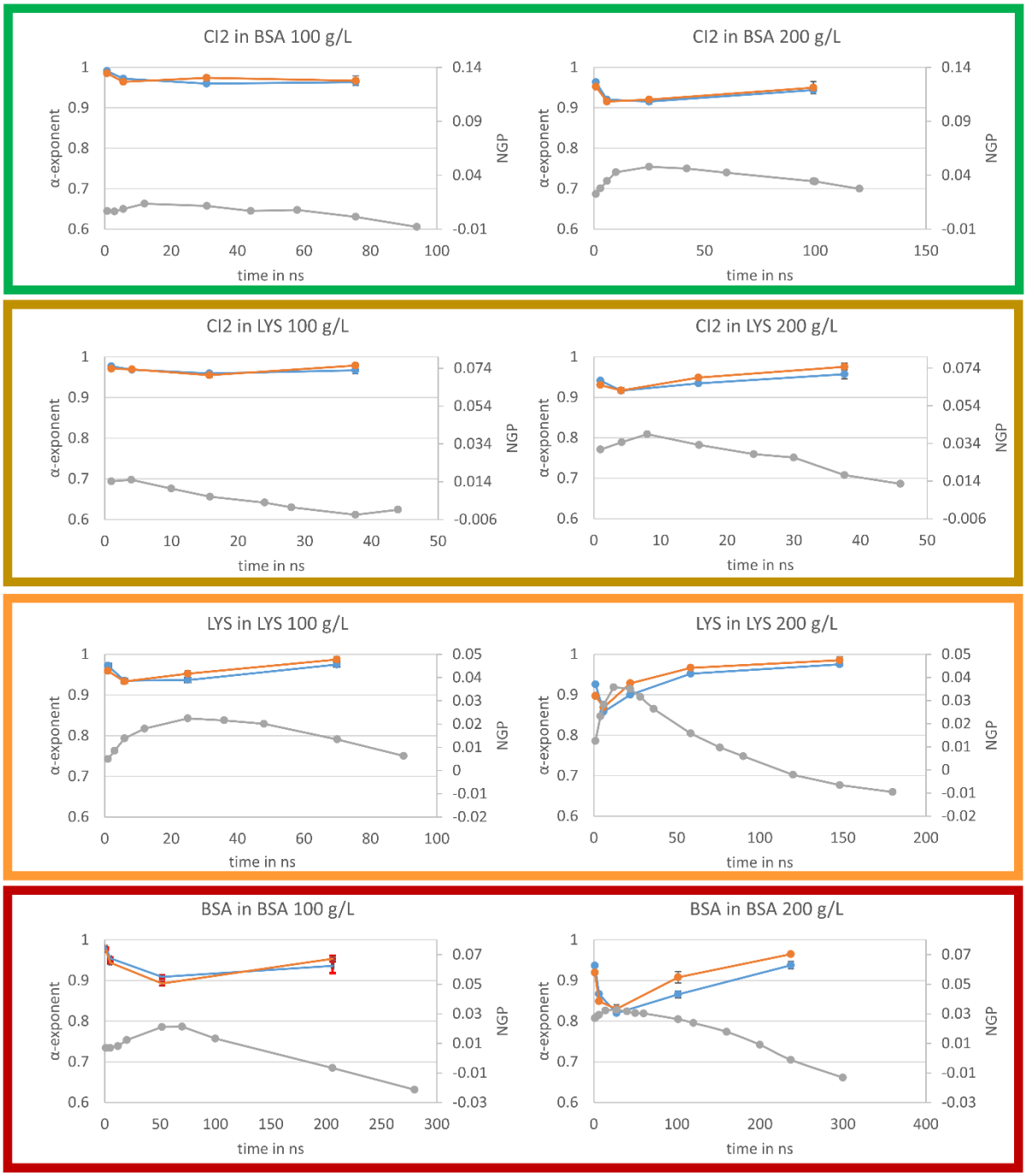

Figure 3.4. Sub-diffusive and non-Gaussianity properties of the crowders and tracer (at concentrations of 100 and 200 g/L of the crowder). The data is represented in the same way as in Figure 3.3.

3.3.2 Cage effects in the protein crowded solutions

The protein dynamics of the above-described crowded systems is consistent with sub-diffusive behaviour arising due to macromolecular crowding. However, the underlying molecular mechanism by which protein crowding causes this phenomenon and its physical origins are not

very well understood. The cage effect hypothesis proposed by previous workers[115] is rigorously tested here. The hypothesis states that macromolecules in a crowded protein solution behave like colloidal particles and exhibit motion akin to rattling in a cage, termed cage effect[132], wherein they are trapped in a transient cage for a finite period of time before ”hopping” to another cage. In contrast to regular Brownian motion, particles do not move freely whilst they are trapped in these cages. Therefore, these particles are expected to exhibit normal diffusion at very short time scales when they are not in close proximity to surrounding particles, but at intermediate time scales these particles would exhibit rattling dynamics, and then exhibit normal Brownian motion at sufficiently long-time scales. In order to quantitatively assess this, Doliwa and Heuer's approach[132] was used to investigate the presence of rattling-in-a-cage type of motion in our simulations. A plot of $\langle x_{12} \rangle$ against $|r_{01}|$ is shown in Figure 3.2C, which was obtained from unfolded trajectories. It is evident from these plots that there is a clear anti-correlation between r_{01} and x_{12}. At higher values of $|r_{01}|$ the plots become noisy because there are very few particles that make very long jumps, reducing the number of data points available for analysis. There is also a higher probability for the particles that make long jumps to exit the transient cage, leading to cessation of the rattling motion.[132] Figure 3.2D shows that, unlike x_{12}, y_{12} does not depend on the magnitude of r_{01}. These findings suggest the presence of a caging effect in crowded protein solutions. The slope of the linear section of the plot is an indicator of the strength of this caging effect. The slope calculated at different 'τ' values in solutions with crowders at concentrations of 50 g/L and 300 g/L is shown in Figure 3.5. As expected, the slope of the tracer CI2 and protein crowders in the 50 g/L solutions was ~0. In the 300 g/L solutions, the slope was initially ~0 but at intermediate time scales the slope was minimum, indicating the existence of a strong cage effect, whilst at longer time scales the slope recovered back to ~0. The x_{12} slopes calculated at intermediate timescales are significantly higher than y_{12} slopes calculated at the same timescales indicating pronounced cage effect as shown in Figure 3.5. These results indicate low cage effect at short timescales followed by maximum cage effect at intermediate timescales and restoration of low cage effect at long timescales. (Figure 3.5)

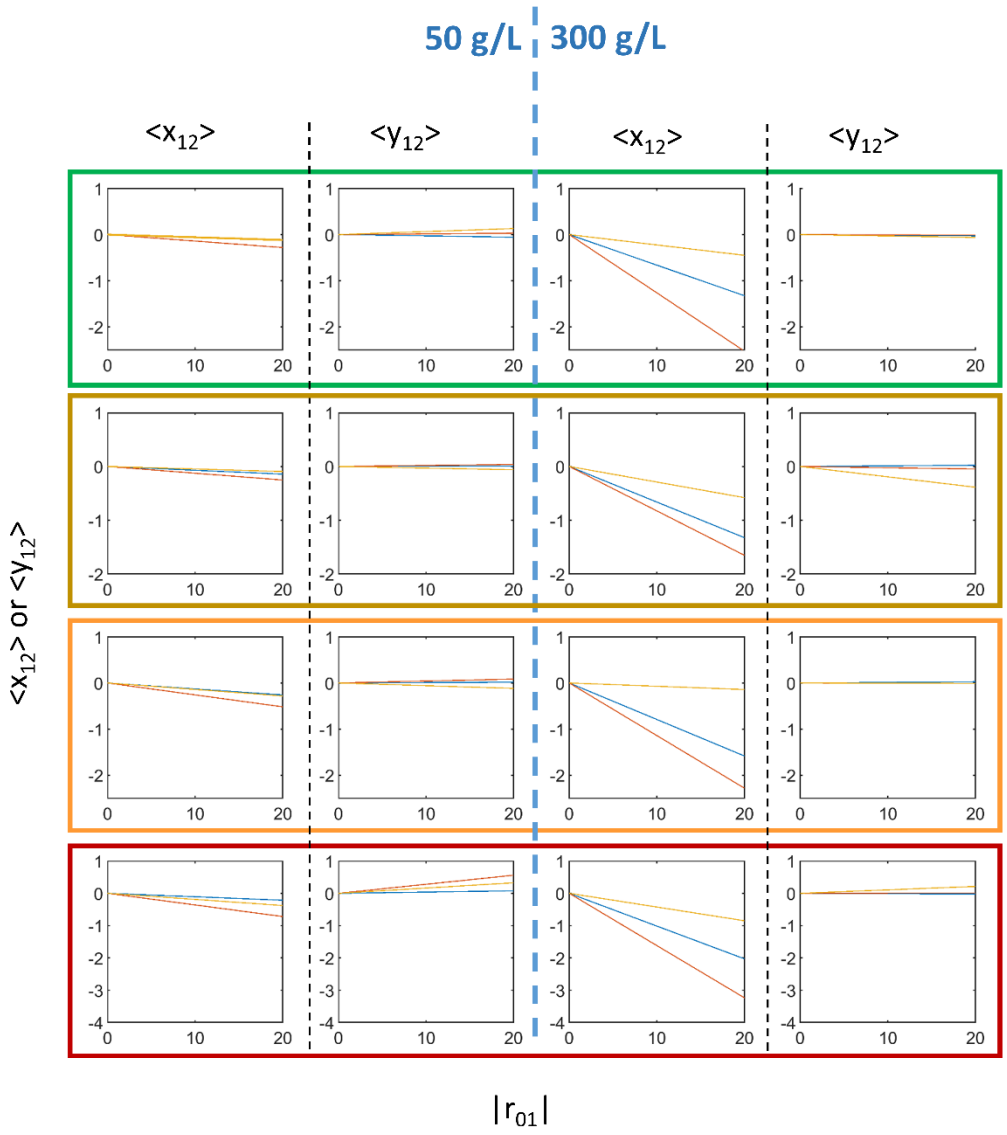

Figure 3.5. Variation of the intensity of cage effects with respect to time and crowder concentration. The straight lines plotted are representative of the slope calculated from the less noisy regions of plots of <x12> or <y12> vs |r01|. The blue, red and yellow lines represent slopes at short, intermediate and long time scales, respectively. The first and second rows highlighted in green and yellow represent the data for the diffusion of CI2 in BSA and lysozyme, respectively. The next two rows highlighted in orange and red represent the data for the self-diffusion of lysozyme and BSA, respectively. The first two columns of every row contain plots of <x12> vs |r01| and <y12> vs |r01| (in that order) at the low protein crowder concentration of 50 g/L. The last two columns contain the same plots at the high protein crowder concentration of 300 g/L.

Weeks and Weitz have shown analytically that the slope of the $|r_{01}|$ vs $\langle x_{12} \rangle$ curve can be used to estimate the value of the α-exponent using the equation below:[131]

$$\alpha_{cage}(\tau) = 1 + \frac{ln\big(1 + slope(\tau)\big)}{ln(2)}$$

Equation 3.2

Using this approach, the α-exponent is calculated from caging effect data for a given lag time τ. The same data at different 'τ' values is obtained by skipping the appropriate number of time frames in a simulation trajectory whilst calculating the displacement vectors. The predicted values of the α-exponent (from the caging effect) of the tracers and crowders in all the simulations (50-300 g/L of both crowders) were calculated and compared with the ones reported in the previous section, as shown in Figure 3.3 and Figure 3.4. The predicted values of the α-exponent are in good agreement with the calculated values for both crowders and tracer under all concentrations of the crowders at all lag times. The consistency in our predictions across different types of proteins with different sizes, net charges and other properties is encouraging. Since sub-diffusive behaviour is the manifestation of multiple mechanisms that do not necessarily constitute anti-correlated displacements[56], the fact that the computed value of the α-exponent obtained from anti-correlated displacements induced by caging effect is consistent confirms the validity of the hypothesis of caging effects causing sub-diffusive behaviour in crowded protein solutions. The same approach described here was earlier used to establish cage effects in the experimental data of protein diffusion on the plasma membrane.[137] However, the cage effect observed in those experiments was in the time scale of a few seconds. The observed anti-correlation of consecutive displacements is similar to the one noted in single particle tracking experiments with dextran crowded solutions, which was explained by fractional Brownian motion.[61]

3.3.3 Non-Gaussianity and ergodicity

In order to probe further the nature of the sub-diffusive behaviour described in the previous section, we investigated the magnitude of deviations from a Gaussian distribution of displacements (Δr) by using a non-Gaussian parameter (NGP, Equation 3.3), in an approach similar to that of previous studies:[138]

$$NGP = \frac{3 < \Delta r^4(\tau) >}{5 < \Delta r^2 >^2} - 1$$

Equation 3.3

The NGP of both crowders and tracers was calculated at all the concentrations and different lag times by choosing appropriate values of τ. It is clear from Figure 3.3 and Figure 3.4 that at low concentrations NGP is very low for both the crowders and tracer with no significant variation with respect to lag time. However, at the highest concentration of 300 g/L there is a clear rise in NGP in all the cases at intermediate time scales. Non-Gaussianity is more prominent around the time scales where anomalous diffusion was identified, as described in the previous sections. More importantly, there is a clear pattern with a slight deviation from Gaussian behaviour at short time scales followed by a rise in non-Gaussianity that eventually reduces at longer time scales. Xue et al. observed increased non-Gaussianity in nanoparticles of comparable size to that of the mesh size of the polymer solution surrounding them.[138] BSA molecules are larger than lysozyme molecules and, therefore, for a given concentration of the crowder, the BSA solution is expected to form larger voids compared to the lysozyme solution. Therefore, the tracer molecule CI2, which is a smaller protein than lysozyme, should exhibit more non-Gaussianity in crowded BSA systems. In line with this argument, the maximum value of NGP for CI2 in a 300 g/L solution of BSA was predicted to be nearly twice as high as that predicted in LYS. The sudden rise in NGP observed in the case of BSA could be due to its larger size, which results in the molecule reaching the cage boundaries in a shorter time. These observations point to a non-Gaussian origin of sub-diffusion, unlike fractional Brownian motion in the case of dextran solutions.[61]

Stochastic processes like fBm are predominantly ergodic in nature, whereas in CTRW deviation from ergodicity have been reported.[56] Whilst investigating the transport of insulin granules inside cells, Tabei et al. used the convergence of TAMSD, which was in turn averaged over the number of particles to infer ergodicity. The authors argued that in an ergodic system, the average TAMSD calculated at a given lag time using simulation trajectories of different lengths should converge once sufficiently long trajectories are chosen.[57] This approach mirrors the way we have assessed convergence in our simulations (Figure 3.6 - 3.9). We chose trajectories of different lengths and calculated diffusion coefficients in all these cases and, for all trajectories beyond certain length, minimal variation in diffusion coefficients was observed.

It can thus be inferred that TAMSD had converged for sufficiently long trajectories, implying ergodicity in our simulation systems.

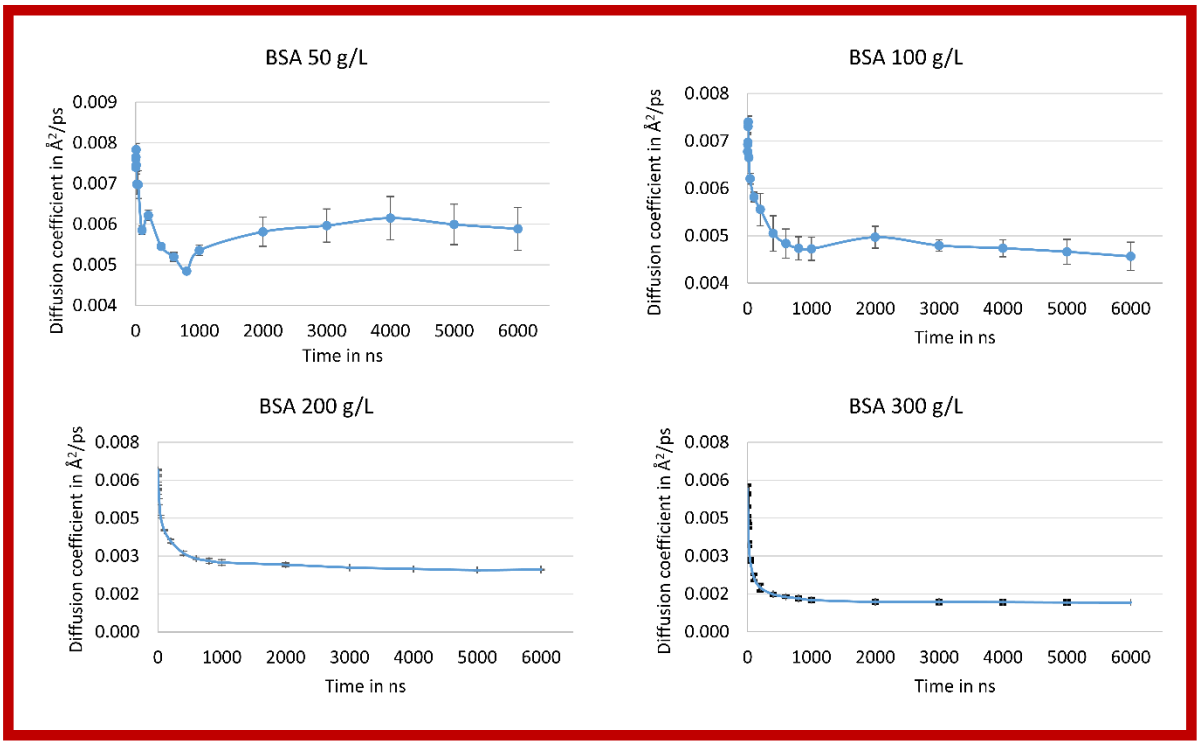

Figure 3.6. Convergence of the diffusion coefficient of BSA at different concentrations in simulations with the full energy term. The error bars represent the standard deviation (n=3).

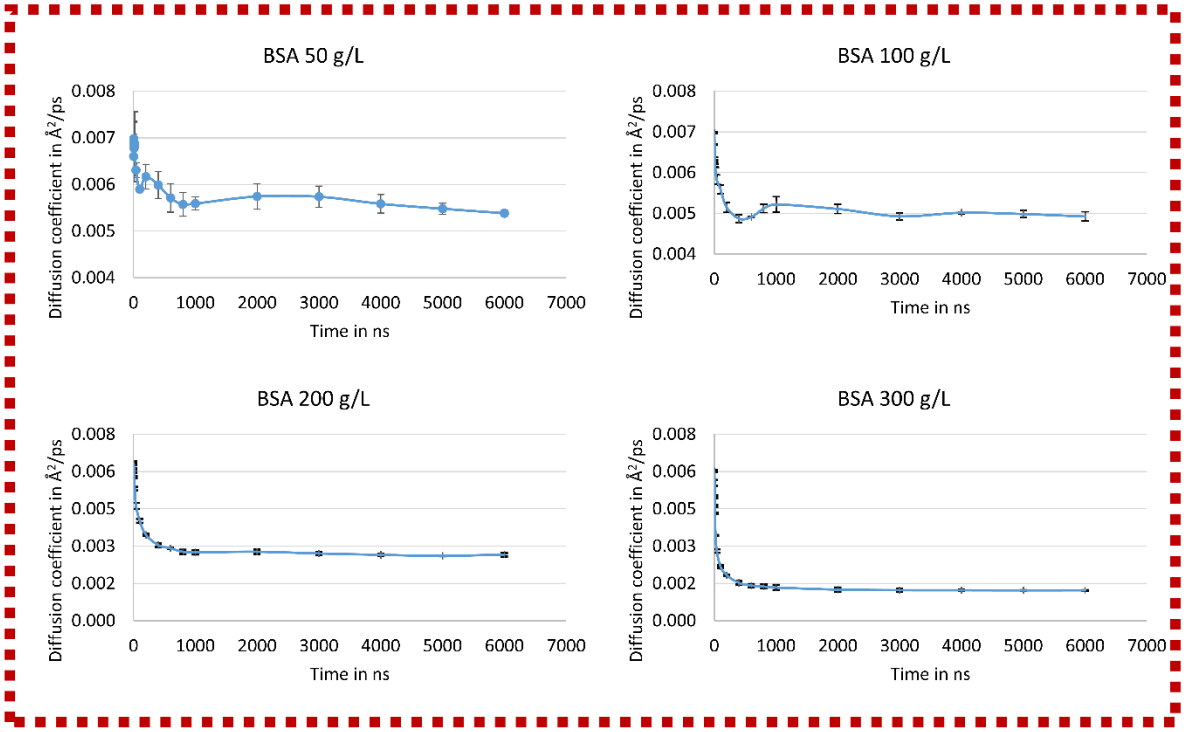

Figure 3.7. Convergence of the diffusion coefficient of BSA at different concentrations in simulations with the soft-core repulsive term only. The error bars represent the standard deviation (n=3).

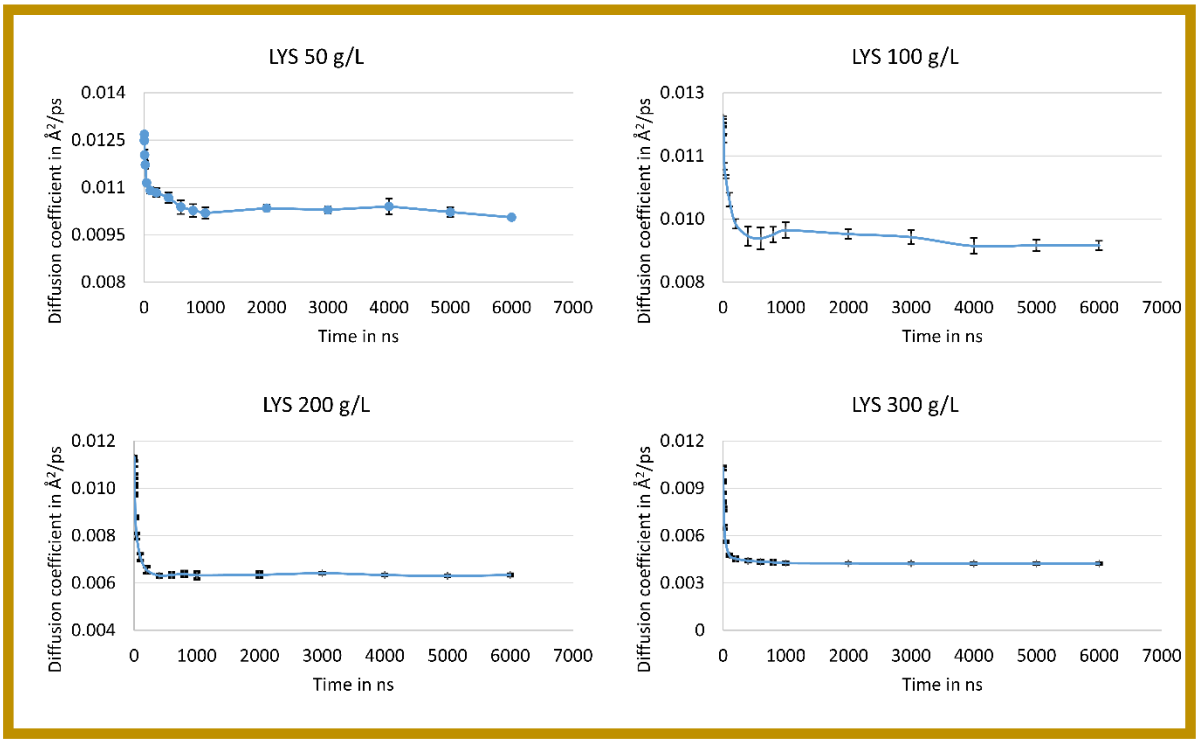

Figure 3.8. Convergence of the diffusion coefficient of lysozyme at different concentrations in simulations with the full energy term. The error bars represent the standard deviation (n=3).

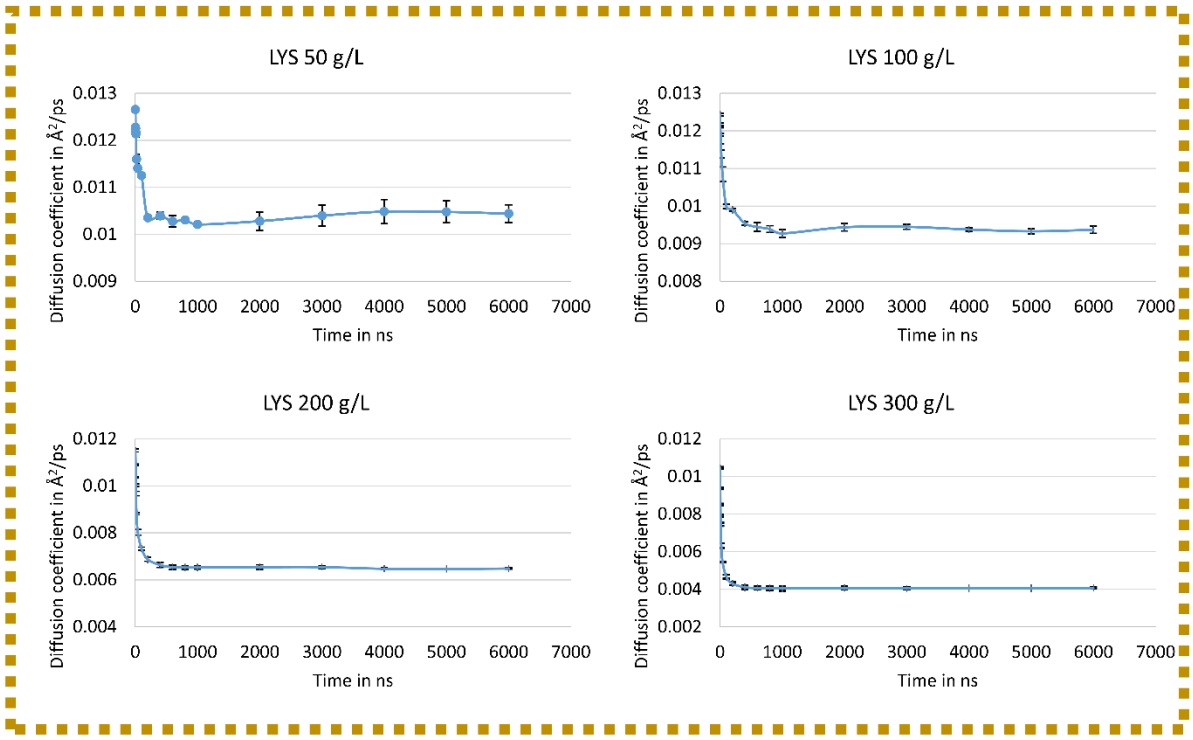

Figure 3.9. Convergence of the diffusion coefficient of lysozyme at different concentrations in simulations with the soft-core repulsive term only. The error bars represent the standard deviation (n=3).

The above findings on ergodicity, non-Gaussianity and anti-correlation show that the behaviour of our simulation systems is similar to that of fBm in finite time intervals, as reported with the numerical simulations of Guggenberger et al.[64] These authors showed that a space-confined particle, whose motion is calculated using a sub-diffusive fBm simulator, initially shows Gaussian behaviour that becomes non-Gaussian at long time scales. This long term non-Gaussianity is attributed to the presence of reflective boundaries. However, in our simulations, at longer time-scales a trend pointing to recovery of Gaussianity is observed. This is due to the fact that, unlike in simulations with a strict reflective boundary, in the case of crowded solutions a particle can cross this boundary at longer time scales and move to a different cage-like structure. Therefore, the movement of a particle in long time scales can be described as being more akin to slow Brownian motion, whereas at intermediate time scales non-Gaussianity due to the reflective nature of cage-like structures is manifested.

3.3.4 Excluded volume effects

Protein molecules in crowded solutions are predicted to form dynamic/transient clusters and exhibit significantly low diffusion rate.[33,112,113] This slow diffusion, especially when the protein molecules form clusters with particularly slow diffusing partners, can potentially be modelled as trapping in CTRW (for a random amount of time), which then gives rise to anomalous diffusion, making cluster formation a possible cause of sub-diffusive behaviour. On the other hand, cluster formation has been proposed as a potential hindrance to caging and, therefore, as reducing anomalous diffusive behaviour.[32,115] However, the role of protein shape and size in regulating sub-diffusive behaviour has not been explored.

The role played by attractive forces between protein molecules in regulating diffusion in timescales of the order of tens of nanoseconds has been previously reported.[113] These studies indicate that the Stokes-Einstein equation is valid in crowded protein solutions, and the slow diffusion of proteins can be explained by the modified Stokes radius as a result of the formation of dynamic clusters.[113] However, it is important to note that the pivotal role played by protein-protein interactions is dependent on the proteins under investigation. Furthermore, given that the time scales of dynamic cluster formation are predicted to be of the order of 1-50 ns[113], the effect of dynamic cluster formation on long-time diffusion coefficients, measured in the microsecond time scale, needs further investigation.

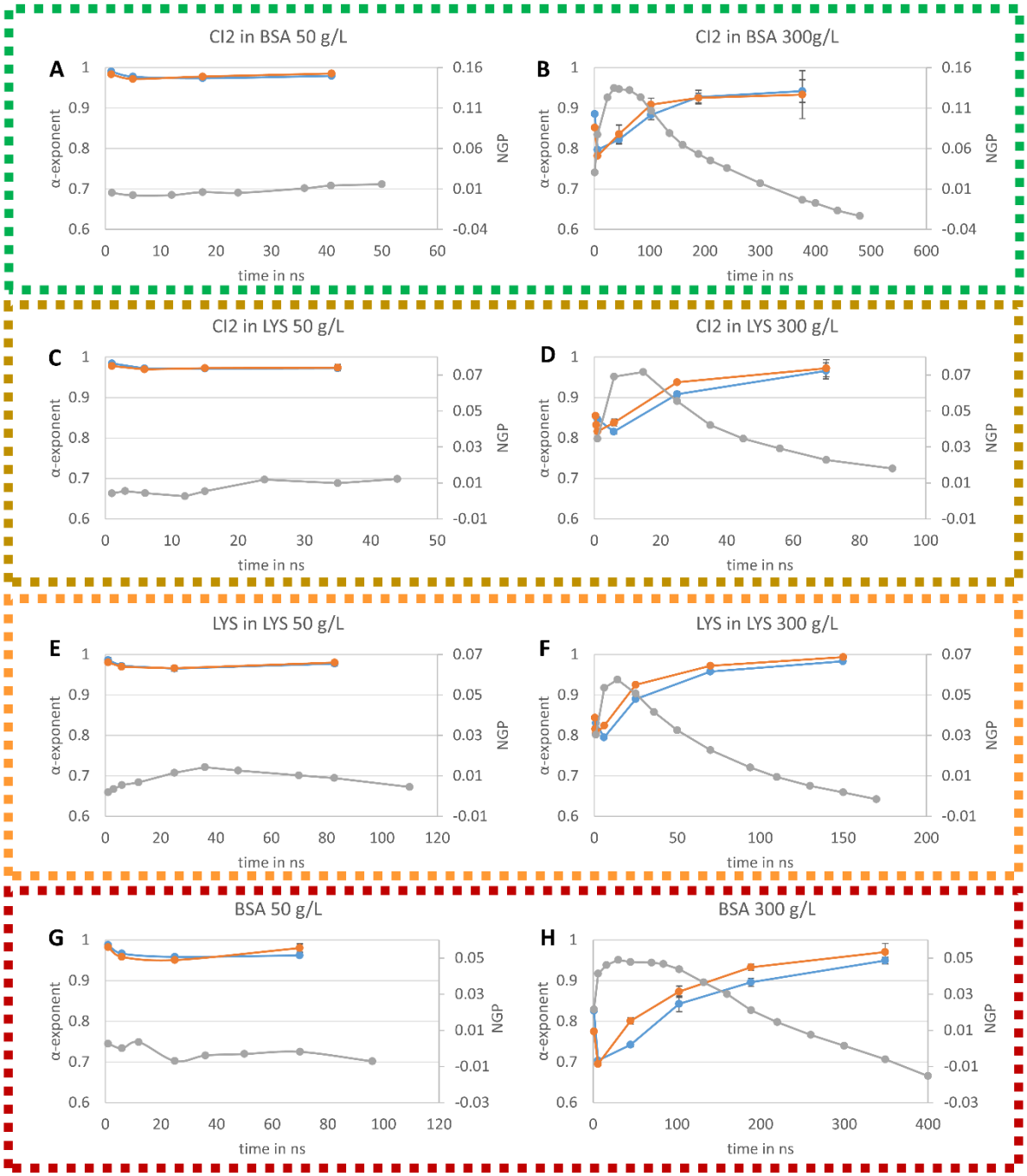

Figure 3.10 The properties of tracer and crowder in the absence of attractive interactions (at concentrations of 50 and 300 g/L of the crowder).The data is represented in the same way as in Figure 3.3. The value of the α-exponent calculated using the log plot and cage effect, and NGP are computed for systems without attractive interactions.

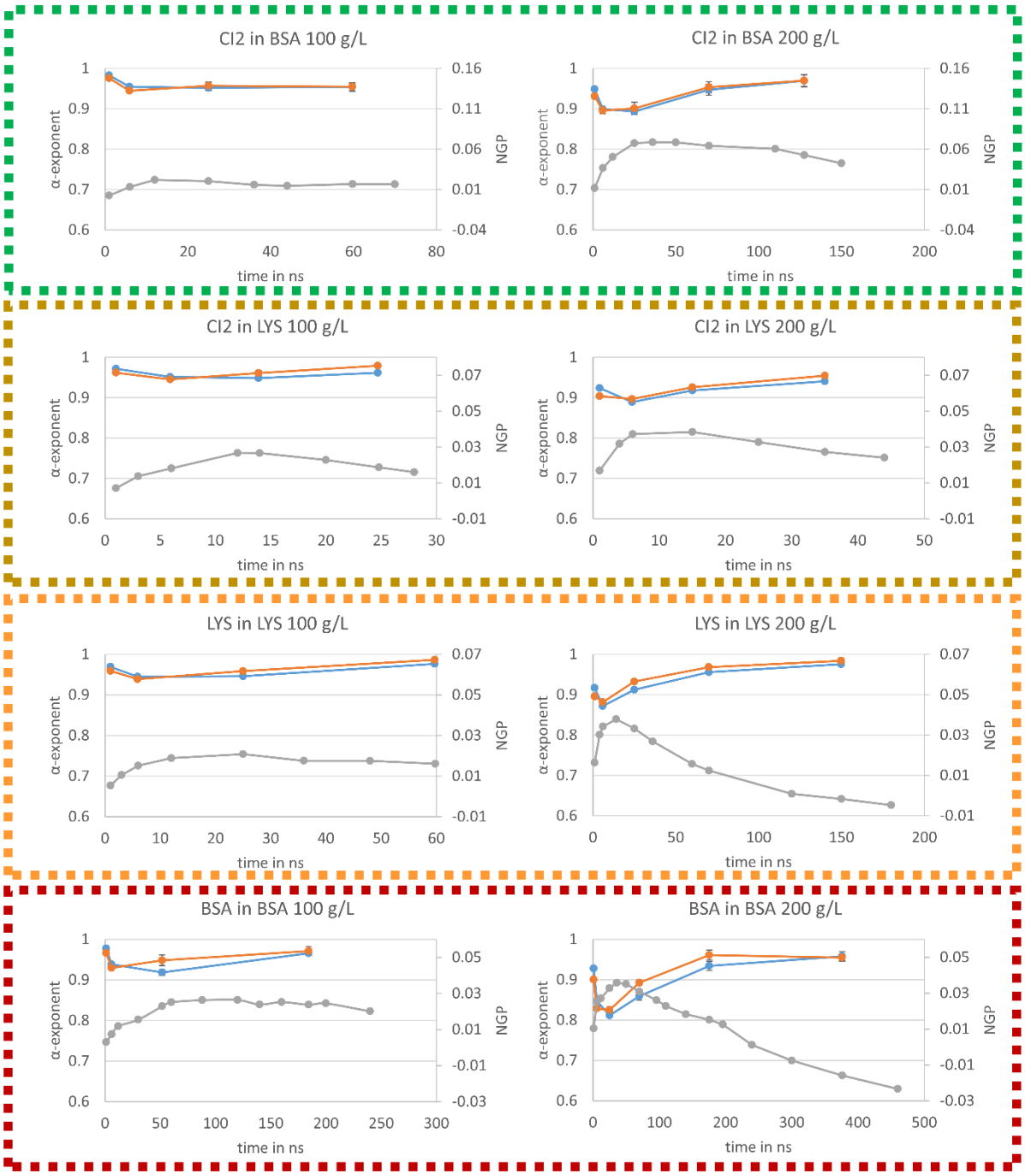

Figure 3.11 The properties of tracer and crowder in the absence of attractive interactions (100 and 200 g/L of the crowder). Error bars represent standard deviation (n=3). The data is represented in the same way as in Figure 3.10.

In order to delineate the effects of cluster formation from those arising from excluded volume, the same set of simulations as described above were conducted using only a soft-core repulsive term to remove attractive interactions between protein molecules. The α-exponent of crowders and tracers was calculated at all concentrations, as shown in Figure 3.10 and Figure 3.11. It can

be seen that sub-diffusion persists despite the lack of attractive interactions. As expected, anti-correlation in the successive displacements due to caging effects is also observed in these simulations with a soft-core repulsive term. At a crowder concentration of 300 g/L, the anomalous diffusion coefficient of BSA reached a minimum value of 0.74 (STD = 0.005) in simulations with the full energy term, and a value of 0.70 (STD = $8x10^{-4}$) in simulations with a soft-core repulsive term, whereas it had a value of 0.80 (STD is 0.001 for both the cases) in the case of LYS in both types of simulation. This is consistent with observations made by Feig and Sugita using all-atom simulations of a single CI2 molecule and eight molecules of BSA/LYS at a concentration of 100 g/L.[32] In their simulations it was shown that BSA has stronger self-interactions than lysozyme does. Therefore, the presence or absence of attractive forces did not significantly affect the α-exponents of LYS ($p = 0.1$). By contrast, due to the relatively stronger interactions between BSA molecules, the absence of attractive forces led to a significant drop in the value of the α-exponent ($p = 0.006$), indicating an increase in sub-diffusive behaviour. In the presence of attractive forces, the value of the α-exponent of CI2 in the crowded environment of LYS was 0.87 (STD = 0.002), indicating minimal sub-diffusion. However, when attractive forces were turned off, the value of the α-exponent reduced ($p = 0.002$) to 0.82 (STD = 0.006). With BSA as a crowder, the value of the α-exponent reduced ($p = 3x10^{-5}$) from 0.83 (STD = 0.002) to 0.80 (STD = 0.002) when attractive forces were turned off. These observations are also consistent with the findings of Feig and Sugita[32], since CI2 interacts more strongly with LYS compared with BSA, and hence there is a larger effect on the value of the α-exponent when attractive forces are turned off. In addition, the value of the α-exponent of BSA was 0.70 and that of LYS was 0.80 in the absence of attractive forces (the significance of this difference was measured using a t-test, $p = 1.5 \times 10^{-6}$). This suggests that caging effects vary between protein species even though neither of them forms clusters. The more pronounced sub-diffusion in BSA in the absence of attractive forces might be due to its large size. Since large-sized crowders can create larger voids in the solution, the probability of protein localization is thus higher. This suggests that the extent of caging effects depends not only on the strength of protein-protein interactions but also on the size of the crowders. Consequently, in systems with the full energy term, overall caging effects are likely to be a function of the basal caging effect (observed in the absence of attractive forces) and the strength of protein-protein interactions. Therefore, caging effects and sub-diffusion are specific to the crowders and tracers present. It is important to emphasise that the maximum caging effect in a given system is observed in the absence of attractive forces. Therefore, sub-diffusion beyond what is predicted from the maximum caging effect must arise from other phenomena. The more

pronounced non-Gaussianity observed in the case of CI2 in BSA and LYS, compared with simulations with the full energy term, could be explained by an increase in excluded volume effects in the absence of attractive forces in the system.

	50 g/L	100 g/L	200 g/L	300 g/L
CI2 in BSA	1.07	1.05	0.98	1.21
CI2 in LYS	1.01	1.02	1.09	0.97
LYS	0.96	0.94	0.98	1.03
BSA	1.10	0.88	0.96	0.95

Table 3.1 Ratios of the long-time diffusion coefficients measured in simulations with only soft-core repulsive interactions in simulations with the full energy term. Dark green colours indicate a high ratio.

The diffusion coefficients calculated using only the soft-core repulsive term (D_{sr}) did not vary significantly from those computed with the full energy term (D_{sim}). The extent of change in the diffusion coefficients when the attractive forces were removed is presented in Table 3.1. Across different concentrations with both crowders the change is nearly 10% in most of the cases, with no specific increasing or decreasing pattern when comparing D_{sr} and D_{sim}. This suggests that the long-term diffusion coefficients measured in the order of one microsecond are largely dependent on excluded volume effects. The decreased diffusion rate of CI2 in the BSA crowded environment in the absence of attractive forces could be due to increased crowding as a result of the increased effective volume occupied by BSA molecules, as described in Figure 3.12. These observations are consistent with those made by Mereghetti et al. on self-crowded solutions myoglobin and haemoglobins.[112] However, the diffusion coefficients of BSA measured in the 0-2 ns time scale in the 300 g/L solutions showed nearly 16% increase when the attractive forces were removed. Diffusion coefficients of LYS in their equally crowded solutions showed a slight decrease in the absence of attractive forces. These findings can be understood by considering the relatively high attractive forces between BSA molecules compared with those between LYS molecules, as noted in the relatively high second order coefficient 'b' in the quadratic function used by Bulow et al.[113] to describe the relation between viscosity and protein volume fraction. The dominance of monomers in LYS solutions under concentrations of less than 0.15% volume fraction (the maximum crowder protein volume fraction in our simulations is 0.135%) has also been experimentally noted.[139] Therefore, in the case of LYS, when weak attractive forces are removed there is a slight increase in the effective radius of the protein, as evidenced by the radial distribution function (RDF) shown in Figure

3.12. This may lead to a minor increase in volume fraction, resulting in a slight decrease in diffusion rates when attractive forces are removed. The short-time diffusion of CI2 in BSA increased by nearly 5% when attractive forces were removed and increased by 10% in LYS. This may be due to the more pronounced interactions between CI2 and LYS, as described above. In summary, our simulations indicate that although the short-time diffusion coefficients in the protein solutions are significantly affected by attractive forces, long-time diffusion coefficients are mostly determined by excluded volume effects.

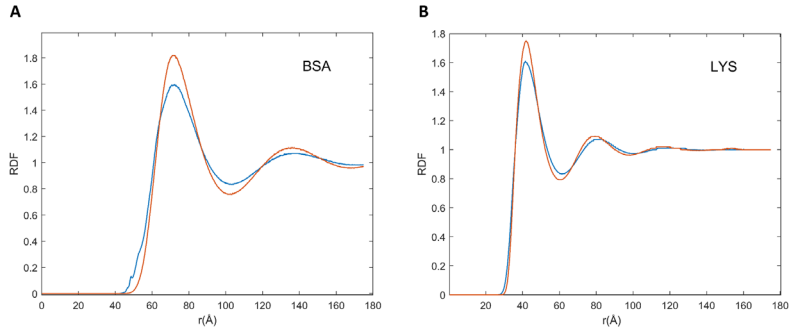

Figure 3.12. Radial distribution functions of BSA and LYS. The red curves correspond to simulations with the soft-core repulsive term only and the blue curves correspond to simulations with the full energy term. The effective radius was approximated as the maximum distance (r) at which RDF~0, and it increases in value by 2.8 Å and 1.6 Å in BSA and LYS respectively, when only the soft-core repulsive term is used. Due to the larger size of BSA compared with LYS, the change in excluded volume due to the small change in the effective radius is more pronounced in the former.

3.4 Discussion and conclusions

Our findings suggest that sub-diffusive behaviour is present in crowded protein solutions and the extent of it depends on the nature of the proteins under consideration. For a given protein solution with a certain crowder species at a given concentration, sub-diffusion mediated by caging effects has a maximum limit. This limit is a function of the proteins under consideration and, therefore, any sub-diffusion stronger than this limit would be the result of phenomena other than caging, such as non-specific interactions mediating sub-diffusion and explained using CTRW models.[140] However, it is evident from the use of a soft-core repulsive energy term only that such non-specific interactions do not play a role in the sub-diffusion observed in our systems, reinforcing the role of caging effects. Recently, it has been shown mathematically that extreme first passage time, the minimum time taken by a searcher in a

group of searchers to reach a target, is lower in the case of sub-diffusive searchers compared with normally diffusing counterparts.[141] This suggests that sub-diffusive behaviour has a vital role to play in biological systems, where molecular encounters drive cellular processes. The implications of caging effects and the subsequent sub-diffusive phenomenon are important in the context of diffusion-limited reactions. Normal diffusion is the underlying assumption made in the derivation of rate constants of diffusion-limited reactions. However, since deviations from normal diffusion are apparent and with varying intensity depending on the protein species and time scales investigated, it is important to account for such deviations using approaches like that of Haugh's[142] especially in the framework of treating biological reaction networks as complex systems. Combining the fact that protein crowded systems emulate the cellular environment[87] and our findings indicating that the strength of sub-diffusion is a result of such crowding, in light of the above mathematical findings, it is possible to infer that cells should maintain crowding for optimal execution of the cellular processes. In fact, such mechanism has already been proposed by Van Den Berg et al. and is termed 'homeocrowding'.[143]

The sub-diffusive behaviour in our systems exhibited features of fractional Brownian motion. However, a more rigorous numerical approach is necessary to establish whether there is fractional Brownian motion in crowded protein systems. It would be interesting to use soft reflective walls that allow particles to escape the confined space in order to explain the restoration of normal diffusion with Gaussian behaviour in long time scales.

The long-time diffusion coefficients in the microsecond timescale appear to be predominantly dependent on excluded volume effects, whilst short-time diffusion coefficients are affected depending on the protein crowder-tracer system. These findings are consistent with the observations of both Mereghetti et al.[112] and Bulow et al.[113] It is also important to note that the simulations reported in this work used rigid protein structures, which could potentially reduce the rates of clustering as a result of their reduced ability to achieve more optimal conformations necessary to maximize protein-protein interactions. However, since the clusters formed have lifetimes of a few tens of nanoseconds[113] one would expect microsecond time scale diffusion coefficients to be affected primarily by excluded volume effects. It is therefore clear that the sub-diffusive and long-time slow diffusive behaviour observed in crowded protein solutions can be explained by volume exclusion. Since the short-time diffusion coefficients and the sub-diffusion observed are dependent on the properties of the proteins (i.e. surface properties such as charge, size and shape), it is important to carefully account for the composition of the

cytoplasmic protein and nucleic acid species when investigating the diffusive behaviour of macromolecules in cell-like environments in these timescales. Finally, this work has shown that SDA can reliably predict the diffusion properties of crowded solutions with more one species of protein.

Chapter 4 Definition of the minimal contents for the molecular simulation of the yeast cytoplasm

4.1 Introduction

Brownian dynamics (BD) simulations performed with SDA can be used to characterise the complex nature of the effects of macromolecular crowding, including the effects on the diffusion of tRNAs. However, a well-defined simulation environment that can accurately predict the crowding effects is necessary to perform such BD simulations. As discussed in chapter 2, there have been many computational studies on the effects of cytoplasmic crowding, in which the crowders are represented at varying levels of detail. All these studies focussed on the effects of prokaryotic cytoplasm, and as a result there have been many attempts in defining a model cytoplasmic environment in prokaryotes. To our knowledge, an equivalent representative definition of the eukaryotic cytoplasm has not been reported in the literature. The key challenges in defining such a simulation cell include identification of the required proteomics datasets and defining appropriate criteria to minimize the size of the simulation cell whilst retaining the properties of the cytoplasmic environment.

In this study, we sought to address the lack of a standard molecular simulation environment for eukaryotes by defining the contents of a simulation cell based on the abundance of proteins, tRNAs and ribosomes in the yeast cytoplasm. A recent yeast proteomics dataset [129] unified abundance data from 21 different datasets, comprising a range of mass spectrometry (MS)-derived datasets, datasets based on green fluorescent protein (GFP)-tagging of yeast proteins and GFP flow cytometry and also a tandem affinity purification (TAP-tagging)-immunoblot dataset. We employed an in-depth proteomics survey of these datasets in order to define a molecular simulation environment for a model eukaryote cell. However, these datasets vary in terms of the growth conditions used to culture the cells, the cellular growth phase, the units in which abundances are reported, and the technique used to measure them. It was therefore necessary to investigate how these factors affect protein abundances reported across the range of datasets. We characterised the internal consistency amongst the datasets and their agreement with other published experimental data, leading to the selection of a proteome composition for the yeast cytoplasmic environment. Consideration of additional experimental data on the macromolecular density and the mass ratio of ribosomal-to-cytoplasmic proteins in the

cytoplasm was also used, allowing the definition of the contents of a molecular simulation cell representative of the yeast cytoplasm.

4.2 Methods

4.2.1 Definition of a eukaryote cell simulation environment

Previous reports of the number of ribosomes in yeast cytoplasm were taken from cell population scale experiments [144] and from cell tomography experiments at single cell level [145], and were compared with the numbers calculated from proteomics datasets. The volume percentage of individual components of the yeast cell were also obtained from cell tomography studies [145], which are in agreement with other cell tomography experiments [146]. Furthermore, we used the recently published unified yeast proteomics dataset that covers a total of 5391 proteins [129].

Proteins associated with the nucleus, cell wall, ribosomes, mitochondria, endoplasmic reticulum and vacuoles were removed from the dataset with the help of GO-slim annotations (http://www.yeastgenome.org/) to assign cellular location to a given protein. Gene ontology analysis of the function of encoded proteins was performed using the webserver Funcassociate 3.0 (http://llama.mshri.on.ca/funcassociate/) [147].

4.2.2 Statistical analysis

The abundances reported for individual ribosomal proteins by any dataset were treated as multiple observations of the number of ribosomes (described in detail in the Results section). Based on this, pairwise statistical two-tailed t-tests for unequal variances between proteomics datasets were performed using an in-house code in MATLAB (https://github.com/BMMG-Curtin/FMOLB) to quantitatively understand the differences and similarities between datasets (Figure 4.3). Where multiple pairwise t-tests were conducted, the Bonferroni correction was applied to address type-I errors, whereby the critical alpha value is divided by the number of pairwise tests. In addition, p-values were adjusted using the Benjamini-Hochberg approach to address type-I errors and the results obtained were found to be qualitatively the same (Figure 4.4 at the end of the chapter). The data was assumed to be normally distributed whilst conducting the above t-tests; therefore, a non-parametric Mann-Whitney U test with the Bonferroni correction was also employed (Figure 4.5 at the end of the chapter). The results of

the *U* test were also found to be qualitatively similar to the results obtained with the t-tests. Pairwise correlations between the functional ontological classes of proteins across different datasets were quantified using the Pearson's correlation coefficient. The Jaccard index was used to quantify the similarities between the ontological profiles obtained for each of the datasets.

4.3 Results

4.3.1 Analysis of internal consistency of yeast proteomics datasets

In order to define the protein composition of a eukaryote molecular simulation cell, the recently published unified yeast proteomics dataset was used [129]. This covers 5391 genes with a total protein mass per yeast cell of 2.7 x 10^{12} Da, which is in good agreement with the total protein mass of a yeast cell previously reported to be 3 x 10^{12} Da [148]. This proteomics dataset comprises data integrated from 21 different datasets, which vary in the type of growth medium used to culture cells, their growth phase and the technique used to measure protein abundances.

The top 200 most abundant proteins were taken from each of the 21 datasets based on their mass (i.e. molecular mass multiplied by their abundance) and were found to account for approximately 70% of the total cytoplasmic protein mass (Figure 4.1). In order to assess the possible influence of cell culture conditions, growth phase and the method used to measure protein abundance on the composition of the yeast cytoplasm, the ontological classes of these proteins were assessed. The systematic names of these proteins were submitted to the Funcassociate 3.0 webserver,[147] which detects over-representation of gene ontologies in a gene list. The number of proteins associated with each gene ontology class was identified for every dataset. Each pair of datasets was then compared by calculating the Pearson's correlation coefficient between the number of proteins associated with each gene ontology class. The Jaccard index was used to quantify the similarities between the sets of gene ontology classes obtained for every dataset. Despite the above differences between the datasets, a similar ontological landscape for the top 200 proteins in each of the datasets was observed, except for one dataset that used N-terminal GFP tagging, YOF [149] (Figure 4.2).

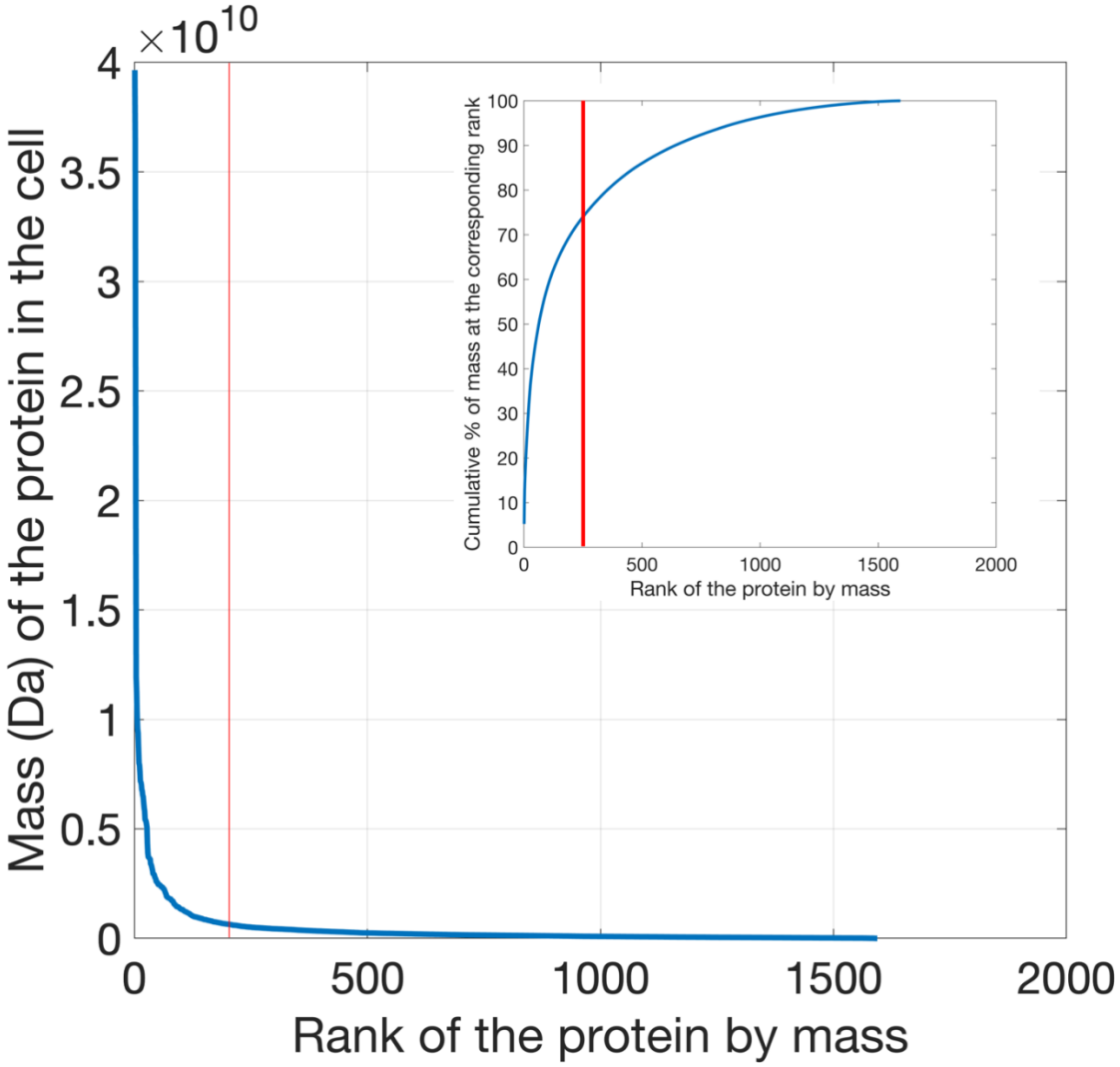

Figure 4.1. Distribution of protein mass (calculated as the product of molecular weight times abundance) per cell plotted as a function of the mass rank of each protein. Proteins in the yeast proteomics dataset were ranked according to their mass, exhibiting a clear exponential decrease as a function of their mass rank in the cell. In the inset the cumulative percentage of mass is plotted as a function of rank. The top 200 cytoplasmic proteins contribute to approximately 70% of the total cell protein mass.

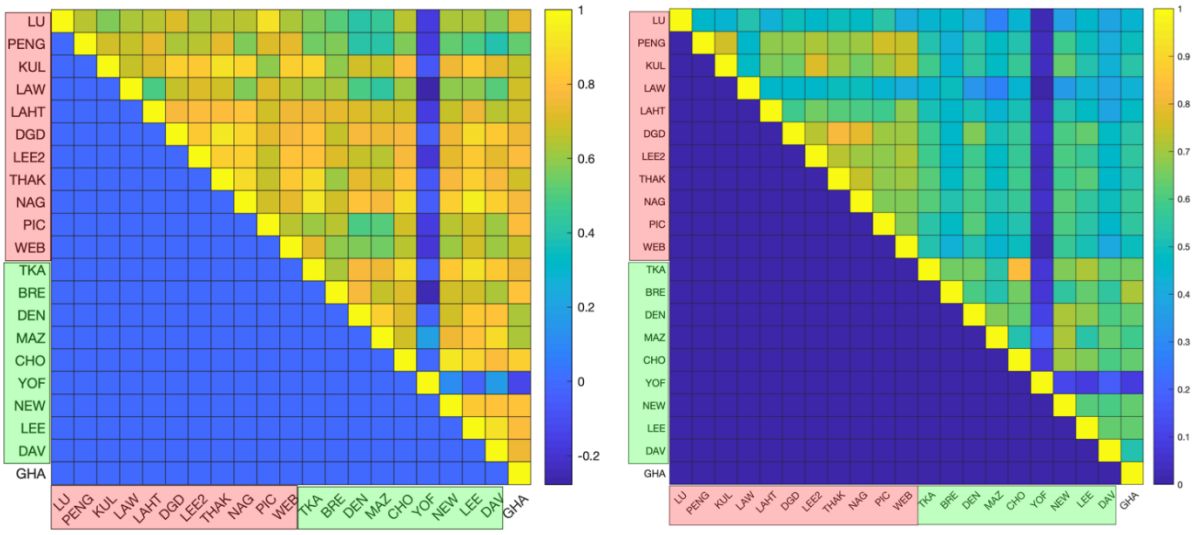

Figure 4.2. Statistical analyses of proteomics datasets. (A) Pairwise correlations between the ontological profiles obtained for the individual datasets. Correlations were measured using the Pearson correlation coefficient, whose values are colour-coded (from the highest correlation in yellow to the lowest correlation in blue). (B) The ontology profile overlap between datasets is quantified using the Jaccard index and the colour-code is the same as in the previous panel. In both panels mass spectrometry based datasets are indicated in red on the axes labelled as LU [150], PENG [151], KUL [152], LAW [153], LAHT [154], DGD [155], PIC [156], LEE2 [157], THAK [158], NAG [159] and WEB [160]; GFP datasets are shown in green on the axes and are labelled as TKA [161], BRE [162], DEN [163], MAZ [164], CHO [165], YOF [149], NEW [166], LEE [167] and DAV [168]; and the TAP-immunoblot dataset is shown in white on the axes and is labelled as GHA [169]. The top 200 proteins are shown to have a similar gene ontology profile across all of the datasets.

Although the gene ontology profiles of the top 200 cytoplasmic proteins are similar across datasets, significant differences in protein abundances were observed. For example, the average coefficient of variation (CV) (measured across the 21 datasets) for the cytoplasmic proteins is 78%. The differences are more marked in the case of ribosomal proteins (CV = 106%).

In order to investigate the internal consistency of the proteomic datasets and their agreement with other published data, ribosomal proteins were examined separately. The protein composition of ribosomes can be assumed to be fixed [170] and there are 79 ribosomal proteins per ribosome. Since the stoichiometry for each ribosomal protein with respect to the ribosome [171] is 1:1, it should be expected that the numbers of each of these ribosomal proteins in a given dataset will lie within a very small range. The identity of the ribosomal proteins was taken from the crystal structure of the eukaryotic ribosome (PDB code 4V88) [172]. The CV of these proteins was computed in every dataset and the average CV of all MS datasets is 69%, whereas the

average CV of GFP datasets is 103%, indicating better internal consistency in MS datasets compared to GFP datasets.

Depending on the consistency between datasets, the numbers reported for a given ribosomal protein across different datasets are expected to vary showing patterns in terms of experimental conditions. In order to test this, the abundances of different ribosomal proteins were compared across different datasets. Given the 1:1 stoichiometry for each ribosomal protein with respect to the ribosome [171], the abundance of each ribosomal protein in each dataset provided an estimate of the number of ribosomes per cell. The average number of ribosomal proteins was therefore calculated to derive an average ribosome per cell value for each dataset. The resulting values were then compared between datasets by performing multiple pairwise t-tests to determine any patterns arising from the growth media, growth phase or the technique used to measure protein abundance (Figure 4.3). High *p*-values were observed in the pairwise tests between the datasets derived from GFP-tagging of proteins, indicating consistency between them. On the other hand, no clear consistency was apparent within the MS datasets, and no patterns were observed that might be accounted for by the growth media or growth phase used during cell culture.

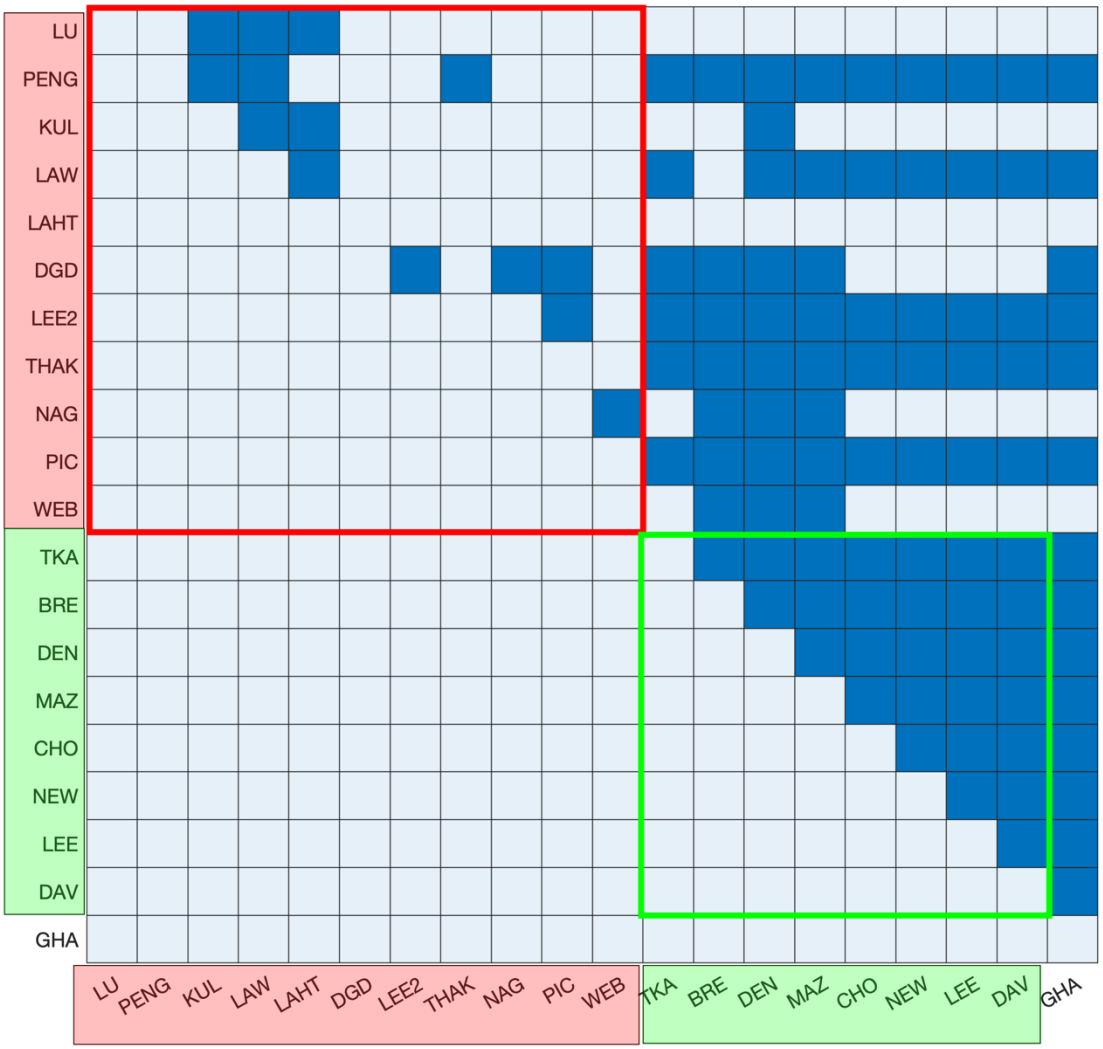

Figure 4.3. Testing of statistical difference between the abundance of ribosomal proteins in each of the datasets. Mass spectrometry-based datasets are shown in red on the axes, GFP datasets are shown in green on the axes and the TAP-immunoblot dataset is shown in white. Ribosomal protein numbers were not reported in the YOF dataset and, therefore, it is not included. The results of t-tests with $p > (0.05/190)$ are coloured dark blue and all others are coloured light blue. GFP datasets exhibit a high level of consistency. There is also consistency among the first five MS datasets. However, there are no discernible patterns in terms of the growth media, growth phase or protein abundance units.

It has previously been reported that there are ribosomal proteins with extra-ribosomal functions in yeast [173]. In order to test if the differences in the abundance (Table 4.1 at the end of the chapter) of ribosomal proteins arise from the fact that some of them perform additional functions and might therefore be produced in excess of the requirements for ribosome synthesis, the mean of means and the mean of medians (across 21 datasets) of ribosomal proteins with extra functions (set I) and other ribosomal proteins (set II) were computed. If

excess production of some ribosomal proteins was due to additional functions, their numbers might be expected to be higher than those of other proteins. However, the mean of means of set I is ~88,400 units, whilst that of set II is ~86,000 units. By contrast, the mean of medians of set I is ~61,700 and that of set II is ~53,157 units. Whilst ribosomal proteins with other functions seem to be abundant, it should be noted that the standard deviations of both sets of proteins are ~25,000. A t-test carried out comparing the means reported for ribosomal proteins in set I and set II has a *p*-value of 0.85 and a similar calculation with medians showed a *p*-value of 0.23. These high p-values suggest that the differences in mean/median abundances do not have statistical significance, suggesting that the differences in the abundances of ribosomal proteins are not due to the extra-ribosomal functions carried out by some of them. The causal relationships of this phenomenon will need to be further investigated.

4.3.2 Selection of datasets

Whilst the gene ontology profiles of the proteomics datasets are similar, they vary widely in the protein abundances reported. The ratio of the median of abundances reported by GFP datasets to the median of MS datasets was calculated for cytoplasmic and ribosomal proteins. We determined that for 74% of cytoplasmic proteins and 84% of ribosomal proteins the medians differ by more than 25%. The differences in the individual protein abundances between the GFP and MS datasets were reported to be possibly due to changes in protein or mRNA stability following GFP tagging [129]. More specifically, in the case of ribosomal proteins, GFP tagging can alter their packing in the ribosome, thereby affecting their turnover dynamics and therefore their abundances [174].

The number of ribosomes, calculated by taking the median of all ribosomal proteins reported in the GFP datasets, revealed an estimated 51,800 ribosomes per cell, whereas previously reported figures are 150,000-300,000 [144] and 169,000-265,000 [145] ribosomes per cell. As discussed earlier, the abundances of ribosomal proteins reported in the GFP datasets are also widely spread, with an average CV of 103%, in contrast to the average CV of 69% in the MS datasets. It was thus decided to omit the GFP datasets from further consideration.

The first five (LU, PENG, KUL, LAW and LAHT) MS datasets report abundances in absolute numbers, whereas the other MS datasets report normalized abundances (with respect to the average of the five MS datasets) [129]. When the median of the first five MS datasets was compared to the median of the other MS datasets individually for every protein, 78% of

cytoplasmic proteins and 96% of ribosomal proteins showed more than 25% difference. These differences may potentially be an artefact of the normalization process. The number of ribosomes inferred from the median abundance of ribosomal proteins of the first five MS datasets was ~130,000, whereas it was only 30,500 when calculated from the other MS datasets. This latter, lower figure is significantly different to previous reports [144,145], as discussed above. The five MS datasets also showed high internal consistency in the pairwise t-tests performed on ribosomal protein abundance compared to the other MS datasets (Figure 4.3). The five MS datasets were originally reported to be highly correlated (with the Pearson correlation coefficient varying from 0.43 to 0.81) [129], which is consistent with our findings. Consequently, it was decided that only the first five MS datasets would be used for the definition of the contents of a molecular simulation cell.

4.3.3 Constraints for the definition of the contents of a simulation cell

A molecular simulation cell should be designed to mimic the environment of the yeast cytoplasm. This requires the inclusion of three important constraints: macromolecular density, the mass ratio of ribosomal-to-cytoplasmic proteins, and the number of ribosomes in the simulation cell.

Macromolecular density is an indirect measure of the excluded volume and, therefore, crowding. The volume of yeast cell has been reported to be 42 μm^3 [175] and from the cell tomography determinations [145] we estimated the cytoplasm in yeast to be 65% of the total cell volume (27.3μm^3). The mass of all the 1374 cytoplasmic proteins in the dataset, excluding ribosomes, was calculated using the mean abundances of all proteins with the above chosen five MS datasets. There are 3 million tRNAs in a yeast cell [144] and, using an average mass of 25,500 Da per tRNA (calculated assuming that there are 75 nucleotides in tRNAs, each weighing an average mass of 340 Da), the total tRNA mass was calculated. The median number of all ribosomal proteins across the five MS datasets was determined to be 126,213, which was used to calculate the ribosomal mass in the yeast cell. The total masses of tRNAs, ribosomes and cytoplasmic proteins was then used to estimate the macromolecular density of the yeast cytoplasm as 90 g/L.

It has been reported that the fractions of ribosomal protein (R-protein), translation protein (T-protein), fixed protein (Q), the proportion of which is independent of growth rate, and

metabolic protein (P-protein), given by, Φ_R, Φ_T, Φ_Q and Φ_P, respectively, are unique for a specific growth rate [34]. Therefore,

$$\Phi_Q + \Phi_P = \frac{\text{Q-Protein}}{A} + \frac{\text{P-Protein}}{A} = C(\textit{growth rate})$$

Equation 4.1

where A is the total protein mass and C is the growth rate specific constant. The total Q- and P-protein content can be divided into cytoplasmic and non-cytoplasmic fractions. Therefore, the previous equation can be rewritten as

$$\Phi_Q + \Phi_P = \frac{\text{non-cytoplasmic}_{(Q+P)}}{A} + \frac{\text{cytoplasmic}_{(Q+P)}}{A} = C(\textit{growth rate})$$

Equation 4.2

$$\frac{\text{non-cytoplasmic}_{(Q+P)}}{A} : \frac{\text{cytoplasmic}_{(Q+P)}}{A} = k(\textit{growth rate})$$

Equation 4.3

The last equation (Equation 4.3) states the assumption that the mass ratio of cytoplasmic to non-cytoplasmic proteins is constant at a given growth rate, from which it follows that cytoplasmic fraction in Q- and P-proteins remains constant. Since the T-protein fraction is a growth rate-dependent constant, the mass ratio of ribosomal-to-total cytoplasmic proteins is constant at a given growth rate. This is the second constraint for the definition of the contents of a simulation cell. The mass ratio of ribosomal-to-cytoplasmic proteins (rib/cyt) was determined to be 0.2229.

The crystal structure of the ribosome is composed of 75 ribosomal proteins [172] and, at such size, it would be computationally challenging to include multiple ribosomes in a single simulation cell. Equally, ignoring the contribution of the ribosome to the excluded volume and macromolecular density would affect the accuracy of a simulation. Therefore, addition of a single ribosome to the simulation cell was decided as the third constraint for the definition of its contents.

4.3.4 Definition of the contents of the simulation cell

The choice of five MS datasets reduced the number of cytoplasmic proteins with abundance data from 1594 to 1374; however, when calculating the macromolecular density of the cytoplasm, data from all 1594 proteins was considered. The total mass of cytoplasmic proteins calculated using abundances in the unified dataset is 7.56×10^{11} Da. The median of the number of molecules reported for a given protein by the five chosen MS datasets was taken as the measure of its abundance in a typical yeast cell. The total mass of a given type of protein was calculated by multiplying its abundance (number of proteins per cell) by its molecular mass, and the protein list was then sorted in descending order of total mass. The top 200 proteins contribute, as mentioned earlier, about 70% of the total cytoplasmic protein mass. The top proteins from the list were chosen due to their significant contribution to the protein mass in the cytoplasm and their abundances were subsequently scaled down to their corresponding value in proportion to only one ribosome (calculated as the abundance 'n' of a protein divided by the 126,213 ribosomes predicted in the MS datasets).

Each of the less abundant cytoplasmic proteins does not contribute significantly to the overall protein mass. However, their collective removal results in a significant loss in protein mass which needs to be accounted for in order to maintain the desired macromolecular density of the simulation cell. Additionally, a number of proteins will contribute to the cytoplasm in fractional units that are lost due to rounding. The number of protein molecules of each of the cytoplasmic proteins was thus multiplied by a scaling factor aimed at maintaining the overall macromolecular density of the simulation cell. The number of protein types was chosen such that their total mass contribution reflects the expected value of the rib/cyt ratio. This was achieved by testing multiple scaling factors under the above-described constraints. Use of a large scaling factor (e.g. 3.0) meant that the rib/cyt ratio could be reached with just 20 different types of proteins, amounting to 119 protein molecules. By contrast, the rib/cyt ratio could not be reached with very low scaling factors (e.g. < 1.8). Although the total number of protein molecules remained in the range 120-130 with all of the scaling factors tested, the observed protein composition was affected significantly with the use of large scaling factors. A range of scaling factors meet the constraints of macromolecular density, rib/cyt ratio and the presence of one ribosome in the simulation cell. However, in order to maintain the most representative composition of cytoplasmic proteins, the lowest possible scaling factor of 1.803 was chosen. This resulted in a final list containing 128 protein molecules belonging to 70 types of proteins (Table 4.2 at the end of the chapter).

Based on the constraint that there should be only one ribosome, the size of the simulation cell was calculated. A total of 126,213 ribosomes are assumed to be present in the cytoplasm, which has a volume of 27.3μm^3. This volume was scaled down to one ribosome unit, which for a cubic simulation cell results in a length of 560 Å. The number of tRNAs was scaled down from 3 million units per cell to the volume of the simulation box, resulting in 22 tRNA units. With one 80S ribosome, 132 protein molecules and 22 tRNAs, the resulting simulation cell has the required total macromolecular density of 90 g/L.

4.4 Discussion

This study shows that the ontological profiles of the most abundant proteins in yeast remains constant despite differences in growth medium and growth phase, indicating that the most abundant proteins constitute the fundamental biochemical framework of the cell. The abundances reported in GFP datasets are affected by tagging, particularly in the case of ribosomal proteins. This has been explained previously on the basis that ribosomal proteins form a compact structure in a single ribosome molecule and the tag attached to them affects their packing. Although this explains the low numbers of ribosomal proteins reported, the cause of the high CV of ribosomal proteins in GFP datasets (CV = 103%), indicating a selective effect of tagging, compared with that of MS datasets (CV = 69%) remains unclear. Moreover, the average number of ribosomes calculated using MS datasets that report abundances in relative units is very low (30,500 units). The causes behind this remain undetermined, although normalization of the data is a possible factor.

Unlike prokaryotic cells, eukaryotic cells have a sophisticated organization of cellular machinery into different organelles with varying macromolecular environments. In order to study the influence of this macromolecular environment, an accurate description of its composition is needed. This was achieved by assigning the cellular location of a protein from its gene annotation data (GO-slim data) and determining the volume percentage of cytoplasm in yeast from cell tomography experiments. The macromolecular density of yeast cytoplasm was found to be 90 g/L, which is three times lower than that of the cytoplasm of *E. coli*. Measurements of the diffusion coefficient of GFP in eukaryotic and prokaryotic cells indicate that the eukaryotic cytoplasm is less crowded,[176] in line with our findings. Crowding in eukaryotic cells is also non-uniform. For example, in the nucleus we have calculated the protein density to be 346 g/L (using the 10-11 volume percentage obtained from cell tomography

experiments [145] and nuclear protein abundances from the dataset [129]). These large macromolecular density differences indicate that an accurate estimate of the macromolecular density of the organelle of interest is necessary.

In conclusion, a simulation cell was defined such that the yeast cellular composition of proteins, the ribosome-to-cytoplasmic protein mass ratio and the macromolecular density are retained. This was achieved by increasing the relative proportion of the most abundant proteins under specific constraints. The resulting simulation cell contains 128 protein molecules belonging to 70 protein types, 22 tRNAs and one 80s ribosome within a cubic cell of 560 Å in length. The simulation cell contents act as a generic representation of the cytoplasm that can be used to study the diffusion and interactions of molecules in the yeast cytoplasmic environment.

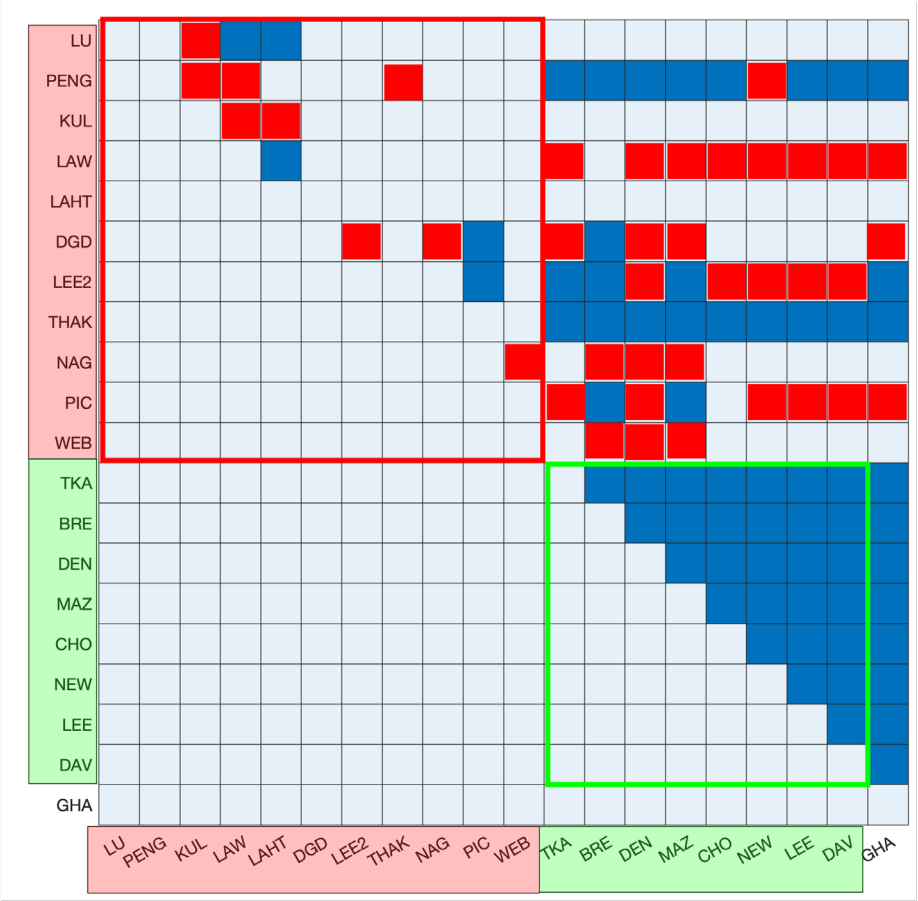

Figure 4.4 Results of the t-tests corrected for type-I errors using the Benjamini-Hochberg approach (an alternative to the Bonferroni correction) with FDR=0.05. Dataset pairs for which p-values > 0.05 are colored in dark blue. Squares in red show deviations from the t-test predictions. The results are qualitatively similar to the t-test predictions and the conclusions drawn from t-tests remain valid.

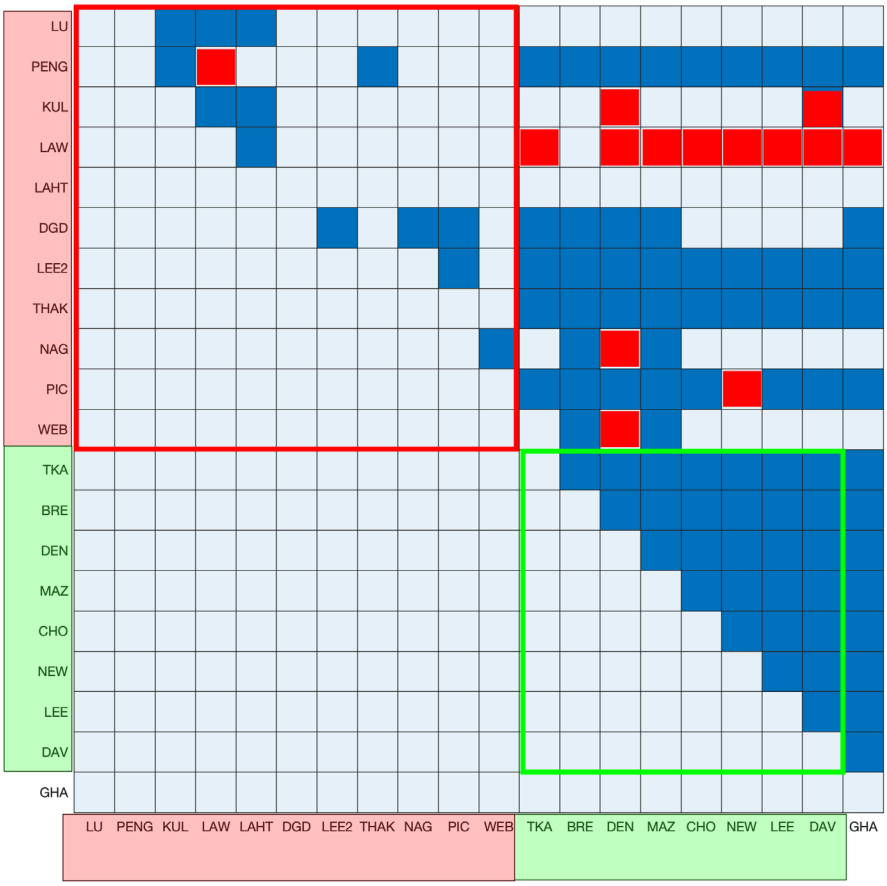

Figure 4.5. Results of the Mann-Whitney U test performed in pairwise manner across the datasets. The Bonferroni correction was applied to address type-I errors. Squares in dark blue show p-values > (0.05/190). Squares in red show the dataset pairs for which the p-values predicted using the Mann-Whitney U test are different from the p-values predicted with the t-tests. The results are qualitatively similar to the t-test results and the conclusions drawn from the t-tests remain valid.

Table 4.1. Mean abundances (averaged over 21 datasets) of ribosomal proteins.

Ribosomal protein	Mean abundance
RPL8A	188027.73
RPP0	130202.95
RPS7A	202553.26
RPS2	152569.09
RPL3	89802.65
RPL5	96389.57
RPS0A	115555.92
RPL4A	81725.33
RPS5	123613.53
RPS4A	103019.95

RPS3	102188.04
RPS1A	88136.51
RPS26A	181591.82
RPL22A	165100.58
RPL7A	80549.23
RPS18A	130570.40
RPL9A	102286.11
RPL10	85701.46
RPL19A	99641.26
RPL31A	162099.18
RPS8A	92966.31
RPS13	118213.17
RPS6A	72525.95
RPL36A	168796.39
RPS9A	81059.75
RPS20	119954.71
RPL18A	79537.21
RPS17A	103341.74
RPL25	103276.00
RPS11A	89415.05
RPS15	94003.66
RPL15A	61326.79
RPS12	92317.14
RPL28	83111.26
RPL2A	50517.13
RPL24A	74903.49
RPL17A	62827.69
RPL14A	80322.78
RPS16A	71760.17
RPL13A	50041.38
RPL21A	61140.15
RPS10A	82701.84
RPL35A	74545.27
RPL20A	50252.89
RPL43A	101502.12
RPS24A	65998.10
RPL11A	51004.52
RPS31	58390.66
RPS14A	68873.54
RPS25A	82158.97
RPL32	66342.15
RPL26A	66121.02
RPS19A	58515.32
RPL27A	59677.32

RPL33A	75015.30
RPL6A	44425.88
RPL30	74292.39
RPL39	126124.28
RPL16A	35233.79
RPS22A	53336.31
RPS21A	80034.68
RPL34A	55699.59
RPS23A	45213.45
RPL38	70481.00
RPL29	80524.73
RPS27A	58361.28
RPL42A	37748.62
RPS28A	57906.56
RPL23A	30148.74
RPL37A	42091.50
RPL41A	95917.02
RPS30A	38135.39
RPL40A	10587.84
RPS29A	18070.34
ASC1	155688.35
STM1*	78257.70

*STM1 is the non-ribosomal protein found in the crystal structure of ribosome.

Table 4.2 The final list of proteins and their structural information. The first column is the rank of the protein when the list is sorted in descending order of mass contributed to the simulation cell. Rows are colour-coded such that green denotes proteins that have an experimentally-determined structure (completely or partially), white denotes proteins that do not have structures but the structures can be predicted using homology modelling, and yellow denotes proteins that do not show sequence similarity to any known structure. There are structures readily available for 34 of the protein types, whilst 32 of the protein types show significant sequence identity with protein structures available and, therefore, their structures can readily be obtained using homology modelling. The remaining 4 types of proteins show no sequence similarity to any structures publicly available and, therefore, *ab initio* modelling approaches can be used to predict their structures.

	Systematic name	Standard name	Description	Number of molecules in simulation cell	PDB ID of the structure / template
1	YBR118W	TEF2	Translational elongation factor EF-1 alpha	13	1F60_A
2	YNL209W	SSB2	Cytoplasmic ATPase that is a ribosome-associated molecular chaperone	5	3GL1_A

3	YDR385W[#]	EFT2	Elongation factor 2 (EF-2), also encoded by EFT1	3	1N0U_A
4	YOL086C	ADH1	Alcohol dehydrogenase	8	4W6Z_A
5	YAL038W	CDC19	Pyruvate kinase	5	1A3W_A
6	YOR133W[#]	EFT1	Elongation factor 2 (EF-2), also encoded by EFT2	3	1N0U_A
7	YLR303W	MET15	O-acetyl homoserine-O-acetyl serine sulfhydrylase	5	2CTZ_A
8	YLR249W	YEF3	Translation elongation factor 3	2	2IWH_A
9	YER091C	MET6	Cobalamin-independent methionine synthase	1	3PPC_A
10	YKR059W	TIF1	Translation initiation factor eIF4A	3	2VSO_A
11	YLR109W	AHP1	Thiol-specific peroxiredoxin	6	4H86_A
12	YMR116C	ASC1	G-protein beta subunit and guanine dissociation inhibitor for Gpa2p	3	3RFG_A
13	YPL106C	SSE1	ATPase component of heat shock protein Hsp90 chaperone complex	1	3C7N_A
14	YAL003W	EFB1	Translation elongation factor 1 beta	4	1IJE_B
15	YPR074C	TKL1	Transketolase	1	1GPU_A
16	YLL039C	UBI4	Ubiquitin	2	4NNJ_B
17	YCL018W	LEU2	Beta-isopropylmalate dehydrogenase (IMDH)	2	3U1H
18	YER043C	SAH1	S-adenosyl-L-homocysteine hydrolase	2	1B3R
19	YDL229W	SSB1	Cytoplasmic ATPase that is a ribosome-associated molecular chaperone	1	3GL1_A
20	YJR109C	CPA2	Large subunit of carbamoyl phosphate synthetase	1	5DOT_A
21	YGL009C	LEU1	Isopropylmalate isomerase	1	4NQY
22	YML028W	TSA1	Thioredoxin peroxidase	3	3SBC_A
23	YDL055C	PSA1	GDP-mannose pyrophosphorylase	2	1TZF_A
24	YPL240C	HSP82	Hsp90 chaperone	1	2CG9_A
25	YLR058C	SHM2	Cytosolic serine hydroxymethyl transferase	1	5Z0Y_A
26	YJL138C	TIF2	Translation initiation factor eIF4A	2	1FUU_A
27	YDR502C	SAM2	S-adenosylmethionine synthetase	1	1O90_A
28	YDR023W	SES1	Cytosolic seryl-tRNA synthetase	1	3QNE_A
29	YHR064C	SSZ1	Hsp70 protein that interacts with Zuo1p	1	5MB9_A
30	YLL050C	COF1	Cofilin	3	1CFY_A
31	YPR145W	ASN1	Asparagine synthetase	1	1CT9_A
32	YCL030C	HIS4	Multifunctional enzyme containing phosphoribosyl-ATP pyrophosphatase	1	5VLB_A
33	YLR180W	SAM1	S-adenosylmethionine synthetase	1	2OBV_A

34	YOR027W	STI1	Hsp90 cochaperone	1	3UQ3_A
35	YHR019C	DED81	Cytosolic asparaginyl-tRNA synthetase	1	5XIX_A
36	YBR025C	OLA1	P-loop ATPase with similarity to human OLA1 and bacterial Ych	1	1NI3_A
37	YMR120C	ADE17	Enzyme of 'de novo' purine biosynthesis	1	1THZ_A
38	YBR126C	TPS1	Synthase subunit of trehalose-6-P synthase/phosphatase complex	1	5HUT_A
39	YGR124W	ASN2	Asparagine synthetase	1	1CT9_A
40	YLR027C	AAT2	Cytosolic aspartate aminotransferase involved in nitrogen metabolism	1	1YAA_A
41	YNL220W	ADE12	Adenylosuccinate synthase	1	5I33_A
42	YHR193C	EGD2	Alpha subunit of the nascent polypeptide-associated complex (NAC)	2	3MCE_A
43	YLR432W	IMD3	Inosine monophosphate dehydrogenase	1	5MCP_A
44	YMR217W	GUA1	GMP synthase	1	5TW7_A
45	YNL138W	SRV2	CAP (cyclase-associated protein	1	1K4Z_A
46	YBR143C	SUP45	Polypeptide release factor (eRF1) in translation termination	1	4CRN_X
47	YLR150W	STM1	Protein required for optimal translation under nutrient stress	1	
48	YKL035W	UGP1	UDP-glucose pyrophosphorylase (UGPase)	1	2I5K_A
49	YLR359W	ADE13	Adenylosuccinate lyase	1	5VKW_A
50	YOL058W	ARG1	Arginosuccinate synthetase	1	1VL2_A
51	YNL064C	YDJ1	Type I HSP40 co-chaperone	1	1NLT
52	YOR184W	SER1	3-phosphoserine aminotransferase	1	6CZY_A
53	YGL105W	ARC1	Protein that binds tRNA and methionyl- and glutamyl-tRNA synthetases	1	4R1J_A
54	YPL037C	EGD1	Subunit beta1 of the nascent polypeptide-associated complex (NAC)	1	NO
55	YKL142W	MRP8	Protein of unknown function; undergoes sumoylation; transcription induced under cell wall stress	1	
56	YDL192W	ARF1	ADP-ribosylation factor	1	5A1U_A
57	YER055C	HIS1	ATP phosphoribosyl transferase	1	2VD3_A
58	YIL041W	GVP36	BAR domain protein	1	
59	YFL045C	SEC53	Phosphomannomutase	1	5UE7_A
60	YEL021W	URA3	Orotidine-5'-phosphate (OMP) decarboxylase	1	3GDK_A
61	YDL137W	ARF2	ADP-ribosylation factor	1	1MR3_F

62	YDR533C	HSP31	Methylglyoxalase that converts methylglyoxal to D-lactate	1	4QYX_A
63	YBR109C	CMD1	Calmodulin	1	6OQQ_B
64	YLR172C	DPH5	Methyltransferase required for synthesis of diphthamide	1	3I4T_A
65	YNL079C	TPM1	Major isoform of tropomyosin	1	
66	YDR071C	PAA1	Polyamine acetyltransferase	1	1B6B_A
67	YGL106W	MLC1	Essential light chain for Myo1p	1	1M45_A
68	YDR177W	UBC1	Ubiquitin-conjugating enzyme	1	1TTE_A
69	YIL138C	TPM2	Minor isoform of tropomyosin	1	5ND5_A
70	YPL225W		may interact with ribosomes, based on co-purification experiments	1	2JYN_A
71	YMR260C	TIF11	Translation initiation factor eIF1A	1	3J80_i

*STM1 is added as a component of ribosome.

[#]These proteins are paralogs encoded by different genes.

Chapter 5 Characterization of the diffusion properties of tRNAs and their complexes in the model yeast cytoplasm

5.1 Introduction

The supply of tRNAs to the translating ribosomes is a crucial aspect of protein synthesis in cells. The tRNAs reach the ribosome as ternary complexes, and this process occurs via their diffusion in the cellular environment. The cellular environment, filled with macromolecules, is known to hinder free diffusion, and the extent of such a crowding effect on the diffusion of tRNAs is not well characterized. The role played by crowding effects in regulating translation can be understood by studying the altered diffusion of tRNAs and their complexes in the presence of macromolecular crowding.

Recent studies on the high-osmolarity glycerol pathway in *Saccharomyces cerevisiae* show that nuclear localization of Hog1p is delayed as a result of severe osmotic stress.[177] Osmotic stress triggers a cascade of reactions resulting in the nuclear localization of Hog1p, which further triggers a transcriptional response to mitigate the osmotic stress effects.[178] Miermont et al. discovered that, although nuclear localization is normal in the presence moderate osmotic stress, severe osmotic stress significantly hinders this process.[177] This behaviour and the associated effects on cell-signalling were attributed to macromolecular crowding arising from the shrinkage of yeast cells.[177] Similarly, Konopka et al. discovered that osmotic stress reduced the diffusion of GFP in *E.coli* by 70-fold, as the volume fraction increased from 0.16 to 0.33.[179] It is thus reasonable to assume that osmotic stress can affect the diffusion of molecules in the translation machinery. However, the magnitude of such an effect is not well understood.

In this part of the study, the effect of macromolecular crowding on the diffusion of tRNAs and ternary complexes was studied using the Brownian dynamics approach employed in Chapter 3. The model yeast cytoplasmic environment described in Chapter 4 was used to construct a suitable simulation system. Additionally, the effect of severe osmotic stress was studied by performing simulations at a higher macromolecular density. These simulations were conducted using a reduced version (containing only the top 4 proteins) of the model cytoplasm. As the high density, reduced system is different from the model cytoplasm at higher density, the

diffusion properties were expected to be affected as a result of the reduced polydispersity. To estimate the extent of such a deviation, similar reduced simulations were conducted maintaining the normal macromolecular density of yeast cytoplasm. The findings of this study suggest that the diffusion of tRNAs and ternary complexes is signicantly affected under the crowded conditions of the yeast cytoplasm. The diffusion is further reduced in the presence of severe osmotic stress, and the tRNAs and their complexes exhibited significant deviations from normal diffusive behaviour.

5.2 Methods

5.2.1 Pre-processing of Ribosome

The structure of the eukaryotic 80s ribosome (PDB ID 4V88[172]) was used for the simulations. Multiple proteins (ribosomal proteins L6, L10, P0, S8, and S17) in the structure contained loops with undetermined coordinates. 25S rRNA, which is expected to contain 3396 bases, has the electron density determined only for residues 3-438, 491-1955, and 2093-3396, whilst 18S rRNA with a total of 1799 total residues has electron density determined only for residues 1-665 and 669-1799. There are also several missing atoms in both the protein and RNA components. As the ribosome is an assembly of multiple protein and RNA chains, modelled loop structures cannot be readily accommodated in the ribosome complex without creating steric clashes. Most of the missing bases in the rRNAs are on the surface of the ribosome, which complicates the process of predicting the right conformations and assembling them onto the ribosome. The missing bases and loops contribute to ~7% of the mass of the ribosome. The effects of the missing bases and loops were not considered in the simulations because (i) the missing mass represents a relatively low percentage, (ii) the diffusion properties investigated are those of ternary complexes and tRNAs and not the ribosome per se, and (iii) it is clear from earlier calculations (Chapter 3) that long-term diffusion coefficients are predominantly affected by excluded volume effects and, therefore, the absent non-specific interactions arising from the missing RNA bases and protein loops likely alter only minimally the diffusion properties of the molecules of interest.

Missing atoms were reconstructed using Swiss PDB viewer[180] and Biovia Discovery Studio (Dassault Systèmes)[181].. Parameters were taken from the CHARMM 36 forcefield[182,183] to generate a PQR file containing coordinates, atomic charges and radii using PDB2PQR[184,185]. To generate the PQR file, hydrogen atoms were added assuming a physiological pH of 7.3.[186]

Electrostatic potential grids were calculated using the parallel focusing option available in APBS[135,187,188] in an approach similar to that used by Baker et al..[189] The grids were calculated at a 150 mM ionic strength[190] and a temperature of 300 K using the linearized Poisson-Boltzmann equation. The final processed grid had a size of 500 x 500 x 500 $Å^3$ and a resolution of 1.0 Å. A modified 2.0 Å version of this grid was used for the calculation of effective charges. These effective charges are the charges placed on specific residues and the termini of the protein molecules, and the phosphate atoms of the RNA molecules such that they replicate the electrostatic grid generated from all the atoms in the molecule. These effective charges are then used in the calculation of forces. The electrostatic desolvation, hydrophobic desolvation, and Lennard-Jones potential grids were calculated at a resolution of 1.0 Å with sizes of 441 x 408 x 413 $Å^3$, 310 x 300 x 300 $Å^3$, and 441 x 408 x 413 $Å^3$, respectively. The Stokes radius of the ribosome was calculated using the 'shell-model from residue level' mode (to improve the calculation time) in HYDROPRO.[28]

5.2.2 Pre-processing of tRNA and ternary complex

The structure of the ternary complex of the phenylalanyl-tRNA of yeast and the elongation factor (EF-Tu) from *Thermus aquaticus* (PDB ID: 1TTT)[191] was used as template to build a homology model of the yeast elongation factor (EF-1α) using Swiss-model[180,192,193]. Upon alignment of the modelled EF-1α and EF-Tu structures (RMSD = 0.279 Å) in the complex, the extent of the conservation of the interactions between the tRNA and EF-Tu identified in the crystal structure of EF-Tu was investigated to ascertain that the structure of this predicted ternary complex was of sufficient quality.

A high degree of conservation of the elongation factor residues involved in the binding of tRNA has been reported across bacteria, plants and animals.[194] The interaction between the amino group of the ester group through which the amino acid is covalently attached to the tRNA and Asn^{285} and His^{273} of EF-Tu was retained in the structure of EF-1α. Similarly, the hydrogen bond between the 2'-OH group of the ester-bond-forming ribose and the conserved Glu^{271} of EF-Tu was also observed in the modelled structure. Glu^{271} in EF-Tu stacks to adenine, which was also observed in the structure of EF-1α. The hydrophobic region formed by Val^{237} and Ile^{231} in EF-Tu was also retained in the modelled structure. The interactions between Lys^{52} of EF-Tu and phosphate groups at positions 74 and 75 of tRNA, the salt bridge between Arg^{300} and the 5'-phosphate, the interactions between ribose and Lys^{90} and Asn^{91}, the co-ordination

of Asp[87] and the phosphates of tRNA bases 3 and 64, the interaction between the main chain of Gly[391] and the riboses of tRNA bases 63 and 64 were all retained in the structure of EF-1α. However, a few interactions observed between EF-Tu and tRNA were not observed in the complex with EF-1α. His[67] in EF-Tu is replaced with Le in EF-1α and, as a result, the interactions between Phe and His[67], and the latter's role in the formation of the ternary complex[191], were not retained in the structure with EF-1α. The interactions between the riboses in tRNA bases 64 and 65 and Gln341 and Thr[350], respectively, were not retained in the structure with EF-1α. Since, most of the interactions observed between the tRNA and EF-Tu [191] were retained in the predicted ternary complex structure of EF-1α, its structure was used for further calculations.

The charge and radius parameters for the modified bases in the tRNA were taken from the CHARMM force field.[183] The protonation states considered in the parametrization of the forcefield correspond to physiological pH[183] and, therefore, the protonation was left unaltered during the generation of the PQR file. The structure of EF-1α was protonated using PDB2PQR[184,185] at a pH of 7.3. The electrostatic and other interaction potential grids were calculated using the same approach described above at a resolution of 1.0 Å and at a size such that the final grids envelope the entire protein.

5.2.3 Pre-processing of Protein crowders

Protein species were classified into four types: (i) proteins that have an experimentally determined structure, (ii) proteins with a high degree of sequence identity (>60%) with structures available in PDB database, (iii) proteins with intermediate sequence identity (<60%) with structures available in the PDB database and (iv) proteins with low sequence similarity with structures available in the PDB database. The first category of proteins were refined further using Swiss-PDB viewer[180] and Biovia Discovery Studio (Dassault Systèmes)[181] and, where necessary, MODELLER[123,124] was used for loop modelling. The homology models of the proteins in the second category with good sequence coverage were built using Swiss-model.[180,192,193] The structures of the proteins with high sequence similarity but poor sequence coverage, as well as proteins in the third category, were built using I-TASSER.[195–197] The structures of the last category of proteins were determined using *ab initio* modelling with QUARK[198,199] and Rosetta (https://www.rosettacommons.org/).

The corresponding PQR files were generated using the same approach described above. All the interaction potential grids were generated at the required sizes, such that the molecules were fully enveloped by the grid.

5.2.4 Determination of the composition of tRNAs

Intracellular tRNAs can be present as charged (amino-acyl tRNAs) and uncharged (free tRNAs) molecules, as well as ternary complexes (aa-tRNA-EF-GTP). About 80% of tRNAs exist in charged form, carrying different amino acids.[200] Hence, 18 out of the 22 tRNAs in the simulation cell were charged.

Charged tRNAs bind to the EF-GTP complex to form a ternary complex with a dissociation constant (K_d) of 3 nM.[201] The concentration of EF determined from the proteomics data (in this case, the five mass spectrometry datasets discussed in Chapter 4), using a yeast cell volume of 42 fL,[175] is 36 µM. It is clear from the observations of Gromadski et al[201] that the concentration of the binary complex of EF and GTP reaches a near saturated value of ~7 nM at a 3 µM concentration of EF. The dissociation constant of the ternary complex can be written as

$$K_d = ([aatRNA] - [TC]).\frac{[EFGTP]}{[TC]}$$

Equation 5.1

where [aatRNA] is the concentration of total charged tRNAs, [TC] is the concentration of ternary complex, and [EFGTP] is the concentration of the binary complex of EF and GTP. The term '[aatRNA]-[TC]' corresponds to the equilibrium concentration of charged tRNAs. Using the above figures, the total number of ternary complex molecules in the simulation cell was determined to be 13. Therefore, of the 22 tRNAs present in the simulation cell, 4 are included as uncharged tRNAs, 5 as free charged tRNAs, and 13 as ternary complex molecules.

5.2.5 Polydispersity index (PDI)

PDI was measured using Grimaldo et al.'s approach[93], and is given by the ratio of the standard deviation to the mean of the Stokes radii of the particles in the simulation cell. PDI is an indicator of the heterogeneity in the size of the crowders.

5.2.6 Simulation system

The processed contents of the yeast cytoplasm were placed in a cubic cell of edge length 560.0 Å using the 'genbox' tool in SDA 7.3.0. 'Genbox' places the molecules inside the simulation cell randomly. The option to place the molecules in the order of their sizes was not chosen to avoid any bias in the resulting configuration. Three initial configurations were generated by randomly placing the molecules in the simulation cell. Given the large size of the ribosome, care was taken to ensure that none of the molecules were placed inside the ribosome. The number of initial configurations was chosen based on the parallelization available in SDA and the computational power required to run these simulations. The simulations were run on multiple machines; Cray XC40 machine (with 24 cores per node with a clock speed of 2.6 GHz), a cluster with 18-core Intel Xeon E5-2695 (Broadwell) series processors with a clock speed of 2.1GHz, and an SGI Linux cluster with 28 cores per node. A simulation runtime of ~0.7 μs per day was noted on all the machines with the 28-core SGI cluster showing the best performance. This could be due to the absence of inter-node parallelization in SDA,[48] and higher number of cores per node in the SGI cluster. The highly crowded systems (described in detail below) showed notably poorer performance clocking ~0.3 μs per day on the 28-core cluster.

Simulations were performed using SDAMM in SDA 7.3.0 using a time step of 0.5 ps, the default temperature (in SDA) of 300 K, and periodic boundary conditions (PBC). The simulations were conducted with just soft-core repulsive forces for 2 μs to remove overlaps between molecules. The simulations were then extended for a further 4.0 μs for equilibration. This was followed by a production run of 22.5 μs. Diffusion coefficients obtained from the time-averaged mean square displacement (MSD) were used to evaluate the convergence of the simulations.

	WHOLE	REDUCED	REDUCED HIGH
Mac.mol. Density (g/L)	90	90	270
Protein types (no.of molecules)	70 (115)	4 (183)	4 (183)
No. of tRNAC	5	9	9
No. of tRNAUC	4	0	0
No. of tRNAEF	13	13	12
No. of ribosomes	1	0	0

Table 5.1 Simulation cell contents of the different types of simulations. Charged tRNAs, uncharged tRNAs and ternary complexes are represented by tRNAC, tRNAUC, and tRNAEF respectively.

Two more systems were set up with reduced cytoplasmic contents to study the effects of osmotic stress, one at a macromolecular concentration corresponding to the yeast cytoplasm (90 g/L) and the other at 270 g/L corresponding to osmotic stress conditions. For these systems the top four protein species from the list described in Chapter 2 (with standard names SSB2, EFT2, ADH1, CDC19) were chosen as crowders and the simulation cell does not contain a ribosome. The concentrations of the crowders were scaled up to match the yeast cytoplasm macromolecular concentration of 90 g/L. The simulations at high concentration were set up by reducing the simulation cell size appropriately (Figure 5.1). The model cytoplasm simulations, simulations with only the top four proteins as crowders at high and normal densities are hereafter referred to as 'WHOLE', 'REDUCED HIGH', and 'REDUCED', respectively. Due to the similarity in their radii of gyration and diffusion properties observed in the WHOLE simulations, charged and uncharged tRNAs were represented as a single species (charged tRNAs) in these simulations to further simplify the simulations. A total of 62 SSB2, 24 EFT2, 59 ADH1, 38 CDC19, 9 free tRNAs and 13 ternary complex molecules (only 12 in the high concentration systems due to rounding) were included in these simulations (Table 5.1). All molecules were prepared using the same approach as in the full cytoplasm simulations. The simulations were set up with three different initial configurations, each in a cubic simulation cell of edge length 560.0 Å (90 g/L) and 388.0 Å (270 g/L). As before, a time step of 0.5ps and a temperature of 300 K was used. These simulations were run for 2.0 μs with soft core repulsive forces only, followed by 4.0 μs (90 g/L) or 10 μs (270 g/L) of equilibration and a production run of 22.0 μs each. Diffusion coefficients were monitored to evaluate convergence. The diffusion coefficient under dilute conditions was calculated using the 'dcc' tool in the SDA 7.3.0 suite. Two-tailed t-test assuming unequal variance was used to calculate p-values.

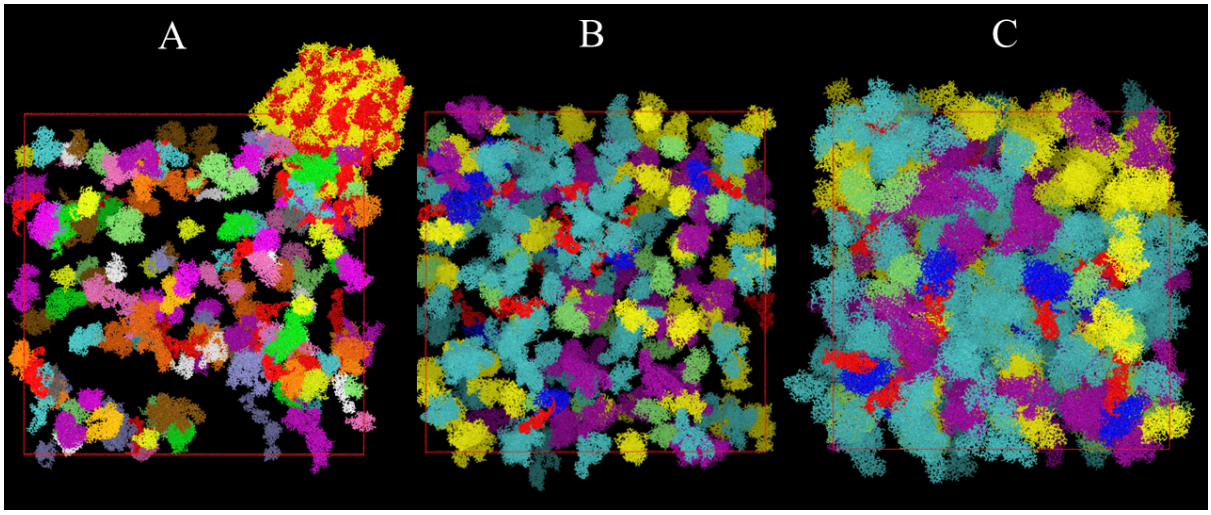

Figure 5.1. Snapshots of the simulation systems of the crowded systems. The RNA molecules in all the systems are shown in red. Red lines show the boundaries of the simulation cell. (A) Whole cytoplasm (WHOLE). The ribosome is the largest molecule in this system, with its proteins shown in yellow and RNA in red. tRNAs are shown in red, elongation factor proteins are shown in blue, and the other crowding proteins are shown in random colours. The edge length of the cubic simulation cell is 560 Å. (B) REDUCED system, where the top four most abundant proteins are the crowders. The macromolecular density of this system is 90 g/L with a simulation cell edge length of 560 Å. Based on the visibly less empty space between the crowders, it can be inferred that the macromolecular density corresponding to the ribosome is distributed evenly across the simulation box. (C) REDUCED HIGH system. The edge length of the cubic simulation cell is 388 Å, with a macromolecular concentration of 270 g/L. Due to the close packing of proteins and the two-dimensional representation of the box, the spaces between molecules are not easily discernible but even distribution of proteins can be assumed.

5.3 Results

5.3.1 Slow diffusion of tRNAs in the crowded cytoplasm

The convergence of predicted diffusion coefficients was evaluated using the same approach described in Chapter 3. Simulation trajectories were truncated at different incremental times and the resulting trajectories were used to calculate diffusion coefficients. The diffusion coefficients were calculated based on the time-averaged mean square displacement (TAMSD) averaged over all of the molecules. The diffusion coefficients computed for each simulation

system ('WHOLE', 'REDUCED' and 'REDUCED HIGH') with different initial configurations were then averaged and plotted as a function of time. Figure 5.2 shows the convergence of the predicted diffusion coefficients across all the systems.

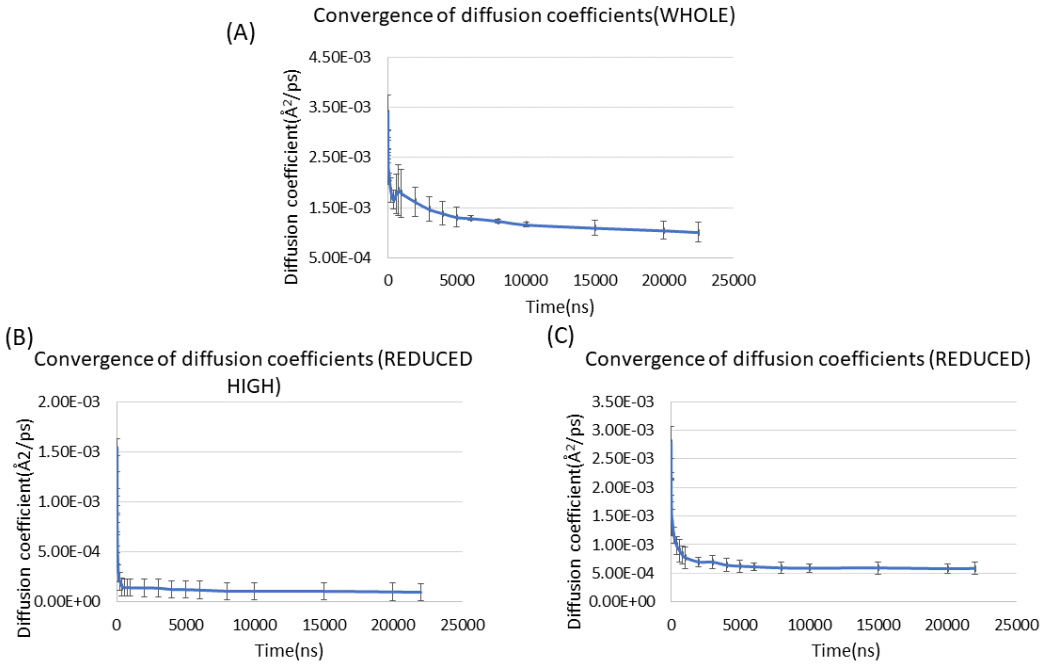

Figure 5.2. Convergence of the predicted diffusion coefficients of the ternary complex of EF-1α with tRNA in the (A) whole cytoplasm (no.of molecules = 13), (B) reduced cytoplasm with high concentration of macromolecules (no.of molecules = 12), and (C) reduced cytoplasm with normal concentration of macromolecules (no.of molecules = 13).

The diffusion coefficients were calculated from production runs totalling 22.5 μs ('WHOLE') and 22.0 μs ('REDUCED HIGH' and 'REDUCED'). The slope of the average TAMSD vs time in the ranges of 0-4500 ns ('WHOLE') and 0-4400 ns ('REDUCED HIGH' and 'REDUCED') were used to calculate the diffusion coefficients of free tRNAs (charged and uncharged) and ternary complexes. The diffusion coefficients reported are the averages over the simulations conducted with different starting configurations.

In the 'WHOLE' system, the diffusion coefficients of charged and uncharged tRNAs were computed to be 1.44 x 10^{-7} cm^2/s and 1.52 x 10^{-7} cm^2/s, respectively (with standard deviations (STDs) of 0.21 x 10^{-7} cm^2/s and 0.75 x 10^{-7} cm^2/s, respectively). The diffusion of the ternary complex was predicted to be slightly slower than that of the free tRNAs, with a diffusion coefficient computed to be 1.01 x 10^{-7} cm^2/s (STD = 0.19 x 10^{-7} cm^2/s). The higher standard deviation in the diffusion coefficients of uncharged tRNAs compared with that of the other

tRNA forms is possibly due to their relatively lower number in the simulation cell. When compared with that of the dilute conditions, the diffusion of ternary complexes, charged and uncharged tRNAs was predicted to be slower by 7.3-, 8.0-, and 7.8-fold, respectively. In comparison, green fluorescent protein (GFP), which has a similar Stokes radius[202], exhibited a 10-fold decrease in diffusion (compared with dilute conditions) in the cytoplasm of *E. coli*[47] (with a macromolecular concentration ~275 g/L). The *in vivo* diffusion coefficient of the free tRNA molecules in the bacterial cytoplasm has been measured to be 0.8×10^{-7} cm^2/s.[30] The slower diffusion rate in bacteria can be understood from the nearly 3-fold higher macromolecular concentration of the bacterial cytoplasm compared to yeast, which is the same increase in macromolecular concentration between the REDUCED HIGH and REDUCED systems. However, it is important to note that the diffusion coefficients predicted from these simulations are within the estimated range of the experimentally observed diffusion coefficients in *E.coli* ($D = 0.8 \times 10^{-7}$ cm^2/s)[30]. It should also be noted that the polydispersity index (PDI) of the model yeast cytoplasm system calculated using the approach of Grimaldo et.al. [93] is 0.38, which is much lower than that of the bacterial cytoplasm simulated by Elcock and McGuffee (PDI=1.05) and Ando and Skolnick (PDI=0.51). This is despite the fact that the number of macromolecular species considered in this study for the representation of the yeast cytoplasm is nearly 50% and 400% more than that of Elcock and McGuffee (50 species) and Ando and Skolnick (15 species), respectively. It is interesting to see that, unlike in prokaryotes, all the most abundant protein species in eukaryotes have a similar size.

The diffusion coefficients of the ternary complex and tRNA(charged) in the 'REDUCED' system are 5.84×10^{-8} cm^2/s and 9.7×10^{-8} cm^2/s, respectively (Table 5.2). The diffusion of the ternary complex in REDUCED system is predicted to be 13-fold slower than that under dilute conditions, whilst the charged tRNAs is predicted to diffuse slower by 12-fold. This corresponds to a 1.7- and 1.5-fold decrease in the diffusion of the ternary complex and tRNAs, respectively, compared with that in the 'WHOLE' system. Although the predicted diffusion coefficients are of the same order of magnitude, when compared with that of the 'WHOLE' system, this disparity points to the importance of defining a well-characterized simulation environment when simulating the diffusion behaviour of proteins under crowded conditions. This difference can be assumed to arise as a consequence of the replacement of the ribosome with an equivalent amount of smaller protein molecules. Since the excluded volume effects are more evenly distributed in the 'REDUCED' system, there are more events where diffusing particles encounter the excluded volume, resulting in a decreased measured diffusion rate.

	Diffusion coefficients (x10^{-7}cm^2/s)		
	WHOLE	REDUCED	REDUCED HIGH
tRNAC	1.44(0.21)	0.97(0.16)	0.16(0.016)
tRNAUC	1.52(0.75)	-	-
tRNAEF	1.01(0.19)	0.58(0.11)	0.1(0.003)

Table 5.2. Predicted diffusion coefficients of tRNA molecules. Standard deviations (n=3) of the data from three initial configurations are shown in the brackets.

As expected, the diffusion of tRNAs and ternary complex molecules is highly reduced in the 'REDUCED HIGH' system. The diffusion coefficients of the ternary complex and tRNA are 9.58 x 10^{-9} cm^2/s and 1.59 x 10^{-8} cm^2/s respectively, which corresponds to a 77- and 73-fold decrease compared with that of dilute conditions. Since the diffusion rate is underestimated in the 'REDUCED' system compared with the 'WHOLE' system, the diffusion of these molecules in an osmotically stressed yeast cell would be expected to be slightly higher than these predictions. These predictions provide nonetheless valuable insights into the diffusion properties of the translation machinery under osmotic stress conditions, and can be readily used to understand the dynamics of protein translation under these conditions by deriving appropriate rate constants using the predicted diffusion coefficients.

5.3.2 Sub-diffusion of tRNA and the EF-1α ternary complex

Sub-diffusion, as defined in terms of the α-exponent, is characterized using a plot of the *log (TAMSD/τ) vs log (τ)*, with the same approach as described in Chapter 3. The α-exponent of tRNAs (both charged and uncharged) and the ternary complex were predicted to have a single consistent slope in the time range of 0-100 ns across all the simulations, followed by a change in slope for the 100-1000 ns regime. (Figure 5.3 and Table 5.3)

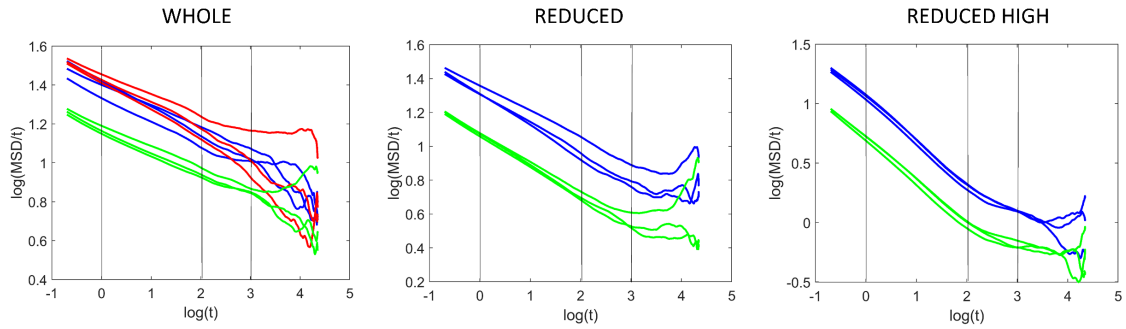

Figure 5.3. Plots of log (MSD/t) vs log (t). The blue-, red- and green-coloured lines represent the data of tRNAC, tRNAUC and tRNAEF, respectively from the three initial configurations of a given system. The ranges between X-axis values 0 and 2, 2 and 3 represented by the black vertical lines show the ranges for which the α-exponent is calculated.

α-exponent						
	0-100 ns			100-1000 ns		
	WHOLE	REDUCED	REDUCED HIGH	WHOLE	REDUCED	REDUCED HIGH
tRNAC	0.87 (9.9×10^{-3})	0.82 (2.4×10^{-2})	0.62 (4.9×10^{-3})	0.91 (2.8×10^{-2})	0.84 (8.9×10^{-3})	0.81 (3.0×10^{-2})
tRNAUC	0.87 (2.0×10^{-2})	-	-	0.86 (6.8×10^{-2})	-	-
tRNAEF	0.89 (1.5×10^{-3})	0.81 (9.3×10^{-3})	0.63 (4.82×10^{-3})	0.91 (2.0×10^{-2})	0.85 (3.1×10^{-2})	0.84 (4.0×10^{-2})

Table 5.3. Predicted alpha exponents of tRNA molecules. Standard deviation (n=3) corresponding to the data from three initital configurations is shown in the brackets.

The α-exponents of charged tRNA, uncharged tRNA, and ternary complex in the 'WHOLE' systems were computed to be 0.87, 0.87, and 0.89, respectively, in 0-100 ns range. These values are close to the α-exponent of BSA (which has a similar Stokes radius) in the solution of CI2 and BSA crowder at a concentration of 100 g/L, which is 0.89. A trend towards normal diffusion was observed in charged tRNA and ternary complex in the range of 100-1000 ns. The α-exponents of the charged tRNA, uncharged tRNA, and ternary complex in this range were computed to be 0.91, 0.86, 0.91, respectively. It should be noted that the high standard deviation in the 100-1000 ns range in all the simulations is a result of the time-averaged approach used to calculate MSD. Interestingly, the sub-diffusion observed in the 'REDUCED' system is more pronounced compared with that of the WHOLE simulations, with an α-exponent of 0.81 (p=0.052) and 0.82 (p=0.005) for the ternary complex and tRNA(charged),

respectively. A similar argument as in the previous section can be invoked to explain this behaviour. Unlike the 'WHOLE' system, in the 'REDUCED' system the mass corresponding to the ribosome is distributed more evenly across the simulation cell, leading to more encounters between molecules, likely enhancing cage-effect like behaviour. A trend towards recovering normal diffusion was observed in this system as well, with an α-exponent of 0.84 and 0.85 for tRNAC and tRNAEF in the 100-1000 ns range.

The most pronounced sub-diffusion across all the systems investigated was observed in the 'REDUCED HIGH' system. The α-exponents of the ternary complex and tRNA (charged) in this system were computed to be 0.63 and 0.62, respectively, in the 0-100 ns range. Since it was shown earlier that the 'REDUCED' system exhibits more pronounced sub-diffusion compared with the 'WHOLE' system, it should be noted that this sub-diffusion might be an overestimate of that in a concentrated form of the 'WHOLE' system. Following the same trend as above, the α-exponent was predicted to be higher in the 100-1000 ns range with a value of 0.84 for the ternary complex and 0.81 for charged tRNA (Table 5.3). Therefore, sub-diffusive behaviour was predicted to persist to a lesser extent in this system in the long-time range. It is interesting to note that the sub-diffusive behaviour of tRNA and its ternary complex is almost the same across all the systems, despite the ternary complex carrying a large protein molecule along with the tRNA. The molecular weight of the EF protein in the ternary complex is ~50 kDa, which is almost the same as that of BSA (molecular weight ~ 60 kDa; Stokes radius = 2.71 nm); however, the Stokes radius of the ternary complex (2.77 nm) is close to that of the tRNA (2.51 nm). This similarity in the Stokes radii might be the reason behind the similarity in their sub-diffusive behaviour.

5.4 Conclusions

The predicted diffusion of tRNA and its complexes was reduced by 7-fold (compared with that of dilute conditions) at a macromolecular concentration of 90 g/L. In contrast, a reduction in the diffusion of GFP by 10-fold required nearly three times the macromolecular concentration in prokaryotic[47] (*E. coli*) cytoplasm conditions. On the other hand, at a macromolecular concentration of 270 g/L in the 'REDUCED HIGH' system, which is nearly the same as that of *E. coli*[47], the diffusion of tRNAs was predicted to decrease by nearly 70-fold. This shows that the composition of the crowded environment along with the macromolecular concentration plays a significant role in regulating diffusion. The same can be inferred from the analysis of

the REDUCED simulations. Although the diffusion coefficients predicted in the REDUCED system are not substantially different from those of the 'WHOLE' system, there is a notable difference.

The sub-diffusive behaviour observed in these systems follows a similar trend as in the crowded solutions of CI2 in BSA or LYS, where short-time sub-diffusive behaviour and a long-time normal diffusive behaviour were observed. However, it is interesting to note that, unlike the simulations described in Chapter 3, the α-exponent remained constant in the range of 0-100 ns, whereas a varying α-exponent was observed in this time-range in the simulations of CI2 in BSA or LYS. This can be due to long-lived cage-effects over relatively long-time scales. However, in agreement with the findings reported in Chapter 3, restoration of normal diffusive behaviour or a trend suggesting the same is observed in these cytoplasm simulations. The role played by non-specific interactions in inducing a CTRW-like behaviour can be ruled out due to the ergodicity observed (inferred from the convergence of the diffusion coefficients in Figure 5.2) in the simulations; however, it would be interesting to verify this using further simulations without attractive forces.

The sub-diffusive behaviour observed in the REDUCED HIGH system reveals that the diffusion of tRNA is far from normal under the high protein concentration conditions of osmotic stress. The role of this sub-diffusive behaviour in regulating the binding of tRNAs to the ribosome needs to be accounted for while studying the protein synthesis dynamics. Although slow diffusion hinders the ability of a molecule to search for its partner, it has been shown analytically[141] that sub-diffusion decreases the extreme first passage time (the minimum time taken for at least one in a group of searchers to reach their target), increasing the frequency of encounters between binding partners. Therefore, the very slow diffusion associated with osmotic stress accompanied by sub-diffusion may, to an extent, lead to compensatory effects.

Due to the reduction in the diffusion coefficients of tRNAs and their complexes in the cytoplasm compared with dilute conditions, protein synthesis dynamics may be affected. Since these effects are more pronounced in cells which are subjected to osmotic stress and, in the case of multicellular organisms, changes in size associated with phenomena like the transport of cells, can contribute to similar effects, the dynamics of translation under these conditions should be analysed considering the associated slow diffusion. Secondly, the sub-diffusion associated with these molecules should be accounted for by using approaches similar to that of Haugh's.[142] In Haugh's approach, anomalous sub-diffusion is considered while calculating the

rate constants of the associated diffusion-controlled reactions.[142] The findings of the current study provide important insights about the diffusion properties of tRNA and ternary complex molecules, which play a pivotal role in characterising the effects of macromolecular crowding on the dynamics of protein translation.

Overall, the study presented here has for the first time established that the diffusion of tRNAs and ternary complexes is slowed down significantly in the crowded environment of the yeast cytoplasm, and that these molecules also exhibit sub-diffusion. In the presence of severe osmotic stress, the diffusion is slowed down further and increased sub-diffusive behaviour is observed. This study establishes a foundation for future mathematical modelling approaches that take into account the slow- and sub-diffusive properties of tRNAs whilst characterizing the function of the translation machinery, which is central to the physiology of all cells.

Chapter 6 Conclusions and future directions

This study investigated the effect of the crowded cytoplasmic environment of yeast on the diffusion properties of tRNAs and ternary complexes. This was achieved by employing a Brownian dynamics simulation approach, using simulation of diffusional association (SDA), that enabled the study of microsecond-time scale diffusion behaviour. The first part of the study focussed on the robustness of this approach in simulating the protein crowded environment with more than one species of protein solute. This was followed by a rigorous analysis of available yeast proteomics datasets to define the contents of a simulation system that mimics the yeast cytoplasmic environment. The model yeast cytoplasmic environment was then characterised using Brownian dynamics simulations.

The robustness of SDA to characterize the slow- and sub-diffusive properties associated with crowded protein solutions with more than one type of protein solute was established by studying the diffusion properties of CI2 in the crowded environment of BSA/LYS. The diffusion coefficients of CI2 predicted using SDA are in good agreement with experiment. There is also good agreement between the sub-diffusion observed in these simulations and the properties revealed in all-atom simulations under the same conditions. Further investigation into the causal relations of this sub-diffusive behaviour revealed that it can be attributed to cage-effects (Chapter 3). Secondly, it was shown that the microsecond time scale slow diffusion in these crowded solutions can be explained by excluded volume effects. The findings of this study also indicate that short-time slow-diffusion and sub-diffusion are crowder-dependent, indicating the importance of a well-defined environment whilst investigating the diffusion properties in cytoplasm-like conditions (Chapter 3).

The model yeast cytoplasmic environment was for the first time defined considering the data from proteomics datasets, cell tomography experiments, and cell population scale experiments (Chapter 4). The simplest model cytoplasm was defined by accommodating only a single ribosome, maintaining the ratio of ribosomal to cytoplasmic protein mass, and the macromolecular density of cytoplasm. The structural information of these contents was either developed or obtained from the literature in order to construct a simulation system. The simulation system cell thus developed can be readily used to study binding and diffusion properties under various conditions, and thus formed the basis for the subsequent study of the effects of diffusion on the translation machinery.

The diffusion properties of tRNA species characterized using the simulation system cell revealed a ~7-fold decrease in the diffusion rate (compared with dilute conditions) of tRNAs and ternary complexes (Chapter 5). The tRNAs also showed mild sub-diffusive behaviour in cytoplasm-like conditions. However, the diffusion properties of tRNAs in a reduced model of the cytoplasm (containing only top 4 proteins) showed clear sub- and slow diffusion. These differences highlighted the importance of defining a model cytoplasm-like environment whilst indicating that the simulations of a reduced model can still provide valuable insights. The properties of tRNAs under osmotic stress were characterized by reproducing hypertonic conditions with a higher macromolecular concentration. These simulations revealed a massive (~70-fold) decrease in the diffusion rate of tRNAs (Chapter 5). However, the associated sub-diffusion might play a compensatory role in the search for synthetases or ribosomes while the tRNA is undergoing aminoacylation or supplying amino acids.[141] In conclusion, the diffusion of tRNAs and ternary complexes is significantly affected in the crowded cytoplasmic environment of the yeast. In comparison, experimental observations of the diffusion of tRNAs in *E.coli*[30] indicate the presence of slower diffusion. This slower diffusion can be explained by a nearly 3-fold higher macromolecular concentration in the cytoplasm of *E.coli,* compared with that of yeast. However, it is important to note that these experiments did not rigorously differentiate charged, uncharged and complexated tRNAs, which also may contribute to the observed differences in the diffusion rates. Secondly, since the sub-diffusion observed in the simulations is more pronounced in the timescales of nanoseconds to microseconds, the experiments, with a time-resolution of milliseconds, could not capture this phenomenon.

In this study, we have successfully demonstrated the capabilities of SDA in simulating 73 different types of proteins, RNAs, and protein-RNA complexes. SDA is currently designed to employ intra-nodal parallelization. However, the future extension of the method for inter-nodal parallelization would enable the execution of simulations on a larger temporal and spatial scale. Such an improved method could provide valuable information like extreme first passage time, diffusion properties over longer timescales and better statistics on the role of non-specific interactions in effecting sub-diffusion. Moreover, recent developments in the field of protein structure prediction can be taken advantage of in pre-processing of the protein crowders. Specifically, the neural network-based structure prediction approach developed by Deepmind[203], Alphafold 2, has been shown to predict structures with a high degree of accuracy[204] with the predicted models scoring above 0.9 on the TM-score (a metric used for assessing the accuracy of the models with a score range of 0 to 1[205]). A highly parallelized

simulation approach in combination with the availability of accurate protein structures plays a key role in understanding the dynamics of protein motion in crowded cell-like environments and the underlying molecular mechanisms, and can also shed light on the organization of processes like translation (as elaborated in the latter paragraphs). It should also be noted that the data from our simulations can also be used to predict the rotational diffusion coefficients, and can provide insights into how they are affected by crowding under both cell-like and protein crowding conditions.

It is important to note that this study has some limitations. Firstly, it does not account for the crowding or sub-diffusion arising due to the cytoskeleton of yeast cells. The diffusion of tRNAs, therefore, can be further limited and there can be sub-diffusion arising as a result of transient binding to the cytoskeleton along with cage-effects. Secondly, the dielectric constant of the solvent in crowded systems is different from that of dilute conditions,[21] and this property is not well characterized for the systems under consideration. Therefore, this effect could not be accounted for in the simulations. Future studies could aim to address these limitations for a better characterization of the simulation conditions.

The experiments conducted by Robbins et al. revealed that protein synthesis decreases by nearly 62% when HeLa cells are exposed to osmotic stress.[206] Macromolecular density in living cells varies in the range of 100-450 g/L,[20] and such a large variation in the macromolecular density should affect protein translation rates. The extent of such an effect across various cell types under different conditions (such as salt concentrations) is not well understood. The simulation approach developed here could be used to estimate these effects by combining the knowledge derived from MD simulations with TASEP approaches.[2,35,39,42] Secondly, the role played by anomalous sub-diffusion, which is more pronounced in systems with higher macromolecular density, in regulating protein translation dynamics is not well understood. This can be incorporated into TASEP models by using a more sophisticated approach, like that of Haugh's[142] for calculating rate constants. The rate constant of a diffusion-controlled reaction is calculated in the Smoluchowski problem by assuming normal diffusion.[142] In Haugh's approach rate constants of diffusion-controlled reactions are calculated without assuming normal diffusive behaviour. However, it is important to note that the sub-diffusive model used in Haugh's approach uses time-varying or space-varying diffusion rate. Therefore, it is important to use an approach that suits the stochastic process underlying sub-diffusion, which appears to be fractional Brownian motion in the case of protein solutions.

Furthermore, the competition between non-cognate, near-cognate and cognate tRNAs for the A-site[207] of the ribosome should also be considered. Along with diffusion properties, the demand-supply dynamics regulate the availability of cognate tRNAs at the A-site.[2] A comprehensive model could be developed that accounts for the complexity of these underlying processes. The differences that may arise whilst comparing the mathematical modelling data with experiments could provide further insights into the channelling mechanism in yeast. In a channelled process, the round-trip of the tRNAs between the ribosomes and aminoacyl-RNA synthetases takes place without the tRNAs entering the bulk of the cytoplasm. Although there is evidence suggesting the existence of such a channelling mechanism in higher organisms, there is not enough data to conclusively prove its presence in lower eukaryotes. The presence of such a channelling mechanism in yeast can be established by calculating translation dynamics in the presence (controlled diffusion of tRNAs between the target molecules) and absence (free diffusion of tRNAs) of channelling, and comparing the results with that of experimental observations. Secondly, in higher eukaryotes, there is evidence for the organization of aminoacyl tRNA synthetases into complexes,[208,209] and the synthetases are found within 40 nm of the A-site during translation.[210] Such localization can result in controlled diffusion of tRNAs between their targets, which may lead to channelling in higher eukaryotes. The manifestation of channelling via such controlled diffusion could be understood better using the knowledge, of the diffusion properties reported in this study. The organization of the translation machinery, like synthetases, at the translation site hints at the possibility of 'translation factories', an analogue of 'transcription factories'[211]. An understanding of the 'channelling' phenomenon developed using the above line of investigation would shed light on the organization of translation into 'translation factories'. The insights provided by a comprehensive mathematical model would also assist in developing a better understanding of disease states like Huntington's disease, which is associated with repeated polynucleotide sequences that iteratively call for a single type of tRNA. The effect of the depletion of these cognate-tRNAs on the overall translation dynamics of the cell could be predicted using such a comprehensive mathematical model.

References

1. Nelson, D. L. & Cox, M. M. *Lehninger Principles of Biochemistry*. (2017).

2. Brackley, C. A., Romano, M. C. & Thiel, M. The dynamics of supply and demand in mRNA translation. *PLoS Comput. Biol.* **7**, e1002203 (2011).

3. Gorochowski, T. E., Ignatova, Z., Bovenberg, R. A. L. & Roubos, J. A. Trade-offs between tRNA abundance and mRNA secondary structure support smoothing of translation elongation rate. *Nucleic Acids Res.* **43**, 3022–3032 (2015).

4. Zhang, G., Hubalewska, M. & Ignatova, Z. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nat. Struct. Mol. Biol.* **16**, 274–280 (2009).

5. Roche, E. D. & Sauer, R. T. SsrA-mediated peptide tagging caused by rare codons and tRNA scarcity. *EMBO J.* **18**, 4579–4589 (1999).

6. Farabaugh, P. J., Zhao, H. & Vimaladithan, A. A novel programed frameshift expresses the POL3 gene of retrotransposon Ty3 of yeast: Frameshifting without tRNA slippage. *Cell* **74**, 93–103 (1993).

7. Mauro, V. P. & Chappell, S. A. A critical analysis of codon optimization in human therapeutics. *Trends Mol. Med.* **20**, 604–613 (2014).

8. Ciandrini, L., Stansfield, I. & Romano, M. C. Ribosome Traffic on mRNAs Maps to Gene Ontology: Genome-wide Quantification of Translation Initiation Rates and Polysome Size Regulation. *PLoS Comput. Biol.* **9**, e1002866 (2013).

9. Adegbuyiro, A., Sedighi, F., Pilkington, A. W., Groover, S. & Legleiter, J. Proteins Containing Expanded Polyglutamine Tracts and Neurodegenerative Disease. *Biochemistry* vol. 56 1199–1217 (2017).

10. Allan Drummond, D. & Wilke, C. O. The evolutionary consequences of erroneous protein synthesis. *Nat. Rev. Genet. 2009 1010* **10**, 715–724 (2009).

11. Wohlgemuth, I., Pohl, C., Mittelstaet, J., Konevega, A. L. & Rodnina, M. V.

Evolutionary optimization of speed and accuracy of decoding on the ribosome. *Philos. Trans. R. Soc. B Biol. Sci.* **366**, 2979–2986 (2011).

12. Lee, J. W. *et al.* Editing-defective tRNA synthetase causes protein misfolding and neurodegeneration. *Nat. 2006 4437107* **443**, 50–55 (2006).

13. Wu, F. & Minteer, S. Krebs cycle metabolon: Structural evidence of substrate channeling revealed by cross-linking and mass spectrometry. *Angew. Chemie - Int. Ed.* **54**, 1851–1854 (2015).

14. Wheeldon, I. *et al.* Substrate channelling as an approach to cascade reactions. *Nat. Chem.* | **8**, (2016).

15. Bauler, P., Huber, G., Leyh, T. & McCammon, J. A. Channeling by Proximity: The Catalytic Advantages of Active Site Colocalization Using Brownian Dynamics. *J. Phys. Chem. Lett.* **1**, 1332–1335 (2010).

16. Negrutskii, B. S. & Deutscher, M. P. Channeling of aminoacyl-tRNA for protein synthesis in vivo. *Proc. Natl. Acad. Sci. U. S. A.* **88**, 4991–4995 (1991).

17. Stapulionis, R. & Deutscher, M. P. A channeled tRNA cycle during mammalian protein synthesis. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 7158–7161 (1995).

18. Barhoom, S. *et al.* Quantitative single cell monitoring of protein synthesis at subcellular resolution using fluorescently labeled tRNA. *Nucleic Acids Res.* **39**, e129 (2011).

19. Cannarrozzi, G. *et al.* A role for codon order in translation dynamics. *Cell* **141**, 355–367 (2010).

20. Feig, M., Yu, I., Wang, P. H., Nawrocki, G. & Sugita, Y. Crowding in Cellular Environments at an Atomistic Level from Computer Simulations. *J. Phys. Chem. B* **121**, 8009–8025 (2017).

21. Nakano, S. I., Miyoshi, D. & Sugimoto, N. Effects of molecular crowding on the structures, interactions, and functions of nucleic acids. *Chem. Rev.* **114**, 2733–2758 (2014).

22. Harada, R., Sugita, Y. & Feig, M. Protein crowding affects hydration structure and dynamics. *J. Am. Chem. Soc.* **134**, 4842–4849 (2012).

23. Asami, K., Hanai, T. & Koizumi, N. Dielectric properties of yeast cells. *J. Membr. Biol.* **28**, 169–180 (1976).

24. Wilcox, A. E., LoConte, M. A. & Slade, K. M. Effects of macromolecular crowding on alcohol dehydrogenase activity are substrate-dependent. *Biochemistry* **55**, 3550–3558 (2016).

25. Ridgway, D. *et al.* Coarse-grained molecular simulation of diffusion and reaction kinetics in a crowded virtual cytoplasm. *Biophys. J.* **94**, 3748–3759 (2008).

26. Ando, T. & Skolnick, J. Crowding and hydrodynamic interactions likely dominate in vivo macromolecular motion. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 18457–18462 (2010).

27. Sharp, S. J., Schaack, J., Cooley, L., Burke, D. J. & Soil, D. Structure and transcription of eukaryotic tRNA gene. *Crit. Rev. Biochem. Mol. Biol.* **19**, 107–144 (1985).

28. Ortega, A., Amorós, D. & García De La Torre, J. Prediction of hydrodynamic and other solution properties of rigid proteins from atomic- and residue-level models. *Biophys. J.* **101**, 892–898 (2011).

29. Carrasco, B. & De La Torre, J. G. Improved hydrodynamic interaction in macromolecular bead models. *J. Chem. Phys.* **111**, 4817 (1999).

30. Plochowietz, A., Farrell, I., Smilansky, Z., Cooperman, B. S. & Kapanidis, A. N. In vivo single-RNA tracking shows that most tRNA diffuses freely in live bacteria. *Nucleic Acids Res.* **45**, 926–937 (2017).

31. Werner, A. Predicting translational diffusion of evolutionary conserved RNA structures by the nucleotide number. *Nucleic Acids Res.* **39**, e17–e17 (2011).

32. Feig, M. & Sugita, Y. Variable interactions between protein crowders and biomolecular solutes are important in understanding cellular crowding. *J. Phys. Chem. B* **116**, 599–605 (2012).

33. Balbo, J., Mereghetti, P., Herten, D. P. & Wade, R. C. The shape of protein crowders is a major determinant of protein diffusion. *Biophys. J.* **104**, 1576–1584 (2013).

34. Klumpp, S., Scott, M., Pedersen, S. & Hwa, T. Molecular crowding limits translation and cell growth. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 16754–16759 (2013).

35. Brackley, C. A., Ciandrini, L. & Carmen Romano, M. Multiple phase transitions in a system of exclusion processes with limited reservoirs of particles and fuel carriers. *J. Stat. Mech. Theory Exp.* **2012**, P03002 (2012).

36. Bonnin, P., Kern, N., Young, N. T., Stansfield, I. & Romano, M. C. *Novel mRNA-specific effects of ribosome drop-off on translation rate and polysome profile*. *PLoS Computational Biology* vol. 13 (2017).

37. Gorgoni, B., Ciandrini, L., McFarland, M. R., Romano, M. C. & Stansfield, I. Identification of the mRNA targets of tRNA-specific regulation using genome-wide simulation of translation. *Nucleic Acids Res.* **44**, 9231–9244 (2016).

38. Ciandrini, L., Stansfield, I. & Romano, M. C. Role of the particle's stepping cycle in an asymmetric exclusion process: A model of mRNA translation. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* **81**, 1–9 (2010).

39. Brackley, C. A., Romano, M. C. & Thiel, M. Slow sites in an exclusion process with limited resources. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* **82**, 1–13 (2010).

40. Zhang, G. *et al.* Global and local depletion of ternary complex limits translational elongation. *Nucleic Acids Res.* **38**, 4778–4787 (2010).

41. Gillespie, D. T. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.* **22**, 403–434 (1976).

42. Greulich, P., Ciandrini, L., Allen, R. J. & Romano, M. C. Mixed population of competing totally asymmetric simple exclusion processes with a shared reservoir of particles. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* **85**, 1–11 (2012).

43. Derrida, B., Domany, E. & Mukamep, D. An Exact Solution of a One-Dimensional Asymmetric Exclusion Model with Open Boundaries. *J. Stat. Phys.* **69**, 1992.

44. Gilchrist, M. A. & Wagner, A. A model of protein translation including codon bias, nonsense errors, and ribosome recycling. *J. Theor. Biol.* **239**, 417–434 (2006).

45. Marshall, E., Stansfield, I. & Romano, M. C. Ribosome recycling induces optimal translation rate at low ribosomal availability. *J. R. Soc. Interface* **11**, 20140589–20140589 (2014).

46. Shah, P., Ding, Y., Niemczyk, M., Kudla, G. & Plotkin, J. B. Rate-limiting steps in yeast protein translation. *Cell* **153**, 1589 (2013).

47. McGuffee, S. R. & Elcock, A. H. Diffusion, crowding & protein stability in a dynamic molecular model of the bacterial cytoplasm. *PLoS Comput. Biol.* **6**, e1000694 (2010).

48. Martinez, M. *et al.* SDA 7: A modular and parallel implementation of the simulation of diffusional association software. *J. Comput. Chem.* **36**, 1631–1645 (2015).

49. Einstein, A. *Investigations on the Theory of the Brownian Movement*. (Dover, 1956).

50. Banks, D. S. & Fradin, C. Anomalous Diffusion of Proteins Due to Molecular Crowding. *Biophys. J.* **89**, 2960–2971 (2005).

51. Weiss, M., Hashimoto, H. & Nilsson, T. Anomalous Protein Diffusion in Living Cells as Seen by Fluorescence Correlation Spectroscopy. *Biophys. J.* **84**, 4043 (2003).

52. Arrio-Dupont, M., Foucault, G., Vacher, M., Devaux, P. F. & Cribier, S. Translational diffusion of globular proteins in the cytoplasm of cultured muscle cells. *Biophys. J.* **78**, 901–907 (2000).

53. Seksek, O., Biwersi, J. & Verkman, A. S. Translational Diffusion of Macromolecule-sized Solutes in Cytoplasm and Nucleus. *J. Cell Biol.* **138**, 131 (1997).

54. Scher, H. & Montroll, E. W. Anomalous transit-time dispersion in amorphous solids. *Phys. Rev. B* **12**, 2455–2477 (1975).

55. Montroll, E. W. Random walks on lattices. III. Calculation of first-passage times with application to exciton trapping on photosynthetic units. *J. Math. Phys.* **10**, 753–765 (1969).

56. Metzler, R., Jeon, J. H., Cherstvy, A. G. & Barkai, E. Anomalous diffusion models and

their properties: Non-stationarity, non-ergodicity, and ageing at the centenary of single particle tracking. *Phys. Chem. Chem. Phys.* **16**, 24128–24164 (2014).

57. Tabei, S. M. A. *et al.* Intracellular transport of insulin granules is a subordinated random walk. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 4911–4916 (2013).

58. Etoc, F. *et al.* Non-specific interactions govern cytosolic diffusion of nanosized objects in mammalian cells. *Nat. Mater. 2018 178* **17**, 740–746 (2018).

59. Mandelbrot, B. B. & Van Ness, J. W. Fractional Brownian Motions, Fractional Noises and Applications. *SIAM Rev.* **10**, 422–437 (1968).

60. Watanabe, C. & Yanagisawa, M. Cell-size confinement effect on protein diffusion in crowded poly(ethylene)glycol solution. *Phys. Chem. Chem. Phys.* **20**, 8842–8847 (2018).

61. Weiss, M. Single-particle tracking data reveal anticorrelated fractional Brownian motion in crowded fluids. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* **88**, 1–4 (2013).

62. Wada, A. H. O. & Vojta, T. Fractional Brownian motion with a reflecting wall. *Phys. Rev. E* **97**, (2018).

63. Basak, S., Sengupta, S. & Chattopadhyay, K. Understanding biochemical processes in the presence of sub-diffusive behavior of biomolecules in solution and living cells. *Biophysical Reviews* vol. 11 851–872 (2019).

64. Guggenberger, T., Pagnini, G., Vojta, T. & Metzler, R. Fractional Brownian motion in a finite interval: Correlations effect depletion or accretion zones of particles near boundaries. *New J. Phys.* **21**, 022002 (2019).

65. Deng, W. & Barkai, E. Ergodic properties of fractional Brownian-Langevin motion. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* **79**, 011112 (2009).

66. Kou, S. C. & Xie, X. S. Generalized langevin equation with fractional gaussian noise: Subdiffusion within a single protein molecule. *Phys. Rev. Lett.* **93**, 180603 (2004).

67. Jeon, J. H., Monne, H. M. S., Javanainen, M. & Metzler, R. Anomalous diffusion of

phospholipids and cholesterols in a lipid bilayer and its origins. *Phys. Rev. Lett.* **109**, 188103 (2012).

68. Wang, B., Anthony, S. M., Sung, C. B. & Granick, S. Anomalous yet Brownian. *Proc. Natl. Acad. Sci.* **106**, 15160–15164 (2009).

69. Chubynsky, M. V. & Slater, G. W. Diffusing diffusivity: A model for anomalous, yet Brownian, diffusion. *Phys. Rev. Lett.* **113**, 098302 (2014).

70. Magdziarz, M., Weron, A., Burnecki, K. & Klafter, J. Fractional brownian motion versus the continuous-time random walk: A simple test for subdiffusive dynamics. *Phys. Rev. Lett.* **103**, 180602 (2009).

71. Meroz, Y., Sokolov, I. M. & Klafter, J. Subdiffusion of mixed origins: When ergodicity and nonergodicity coexist. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* **81**, 010101 (2010).

72. Axelrod, D., Koppel, D. E., Schlessinger, J., Elson, E. & Webb, W. W. Mobility measurement by analysis of fluorescence photobleaching recovery kinetics. *Biophys. J.* **16**, 1055 (1976).

73. Kang, M., Day, C. A., Kenworthy, A. K. & DiBenedetto, E. Simplified equation to extract diffusion coefficients from confocal FRAP data. *Traffic* **13**, 1589 (2012).

74. Periasamy, N. & Verkman, A. S. Analysis of Fluorophore Diffusion by Continuous Distributions of Diffusion Coefficients: Application to Photobleaching Measurements of Multicomponent and Anomalous Diffusion. *Biophys. J.* **75**, 557–567 (1998).

75. Dey, D. *et al.* Line-FRAP, A Versatile Method to Measure Diffusion Rates In Vitro and In Vivo. *J. Mol. Biol.* **433**, 166898 (2021).

76. Kang, M. & Kenworthy, A. K. Complex Applications of Simple FRAP on Membranes. in *Biomembrane Frontiers: Nanostructures, Models, and the Design of Life* (eds. Faller, R., Longo, M. L., Risbud, S. H. & Jue, T.) 187–221 (Humana Press, 2009). doi:10.1007/978-1-60761-314-5_8.

77. Schlessinger, J. *et al. Lateral transport on cell membranes: Mobility of concanavalin A receptors on myoblasts (receptor mobility/membrane diffusion/photobleaching*

recovery). *Cell Biology* vol. 73 (1976).

78. Webb, W. W., Barak, L., Tank, D. W. & Wu, E. S. Molecular mobility on the cell surface. *Biochem. Soc. Symp.* **No. 46**, 191–205 (1981).

79. Kitamura, A. & Kinjo, M. Determination of diffusion coefficients in live cells using fluorescence recovery after photobleaching with wide-field fluorescence microscopy. *Biophys. Physicobiology* **15**, 1 (2018).

80. Phair, R. D. & Misteli, T. High mobility of proteins in the mammalian cell nucleus. *Nat. 2000 4046778* **404**, 604–609 (2000).

81. Kitamura, A., Nakayama, Y. & Kinjo, M. Efficient and dynamic nuclear localization of green fluorescent protein via RNA binding. *Biochem. Biophys. Res. Commun.* **463**, 401–406 (2015).

82. Luby-Phelps, K., Castle, P. E., Taylor, D. L. & Lanni, F. Hindered diffusion of inert tracer particles in the cytoplasm of mouse 3T3 cells. *Proc. Natl. Acad. Sci. U. S. A.* **84**, 4910 (1987).

83. Chen, H., Farkas, E. R. & Webb, W. W. *Chapter 1 In Vivo Applications of Fluorescence Correlation Spectroscopy*. *Methods in Cell Biology* vol. 89 (Elsevier Inc., 2008).

84. Ramadurai, S. *et al.* Lateral diffusion of membrane proteins. *J. Am. Chem. Soc.* **131**, 12650–12656 (2009).

85. Bacia, K., Kim, S. A. & Schwille, P. Fluorescence cross-correlation spectroscopy in living cells. *Nat. Methods 2006 32* **3**, 83–89 (2006).

86. Roos, M. *et al.* Coupling and Decoupling of Rotational and Translational Diffusion of Proteins under Crowding Conditions. *J. Am. Chem. Soc.* **138**, 10365–10372 (2016).

87. Wang, Y., Li, C. & Pielak, G. J. Effects of proteins on protein diffusion. *J. Am. Chem. Soc.* **132**, 9392–9397 (2010).

88. Pagès, G., Gilard, V., Martino, R. & Malet-Martino, M. Pulsed-field gradient nuclear magnetic resonance measurements (PFG NMR) for diffusion ordered spectroscopy

(DOSY) mapping. *Analyst* **142**, 3771–3796 (2017).

89. Kuchel, P. W. *et al.* Stejskal–tanner equation derived in full. *Concepts Magn. Reson. Part A* **40A**, 205–214 (2012).

90. Sinnaeve, D. The Stejskal–Tanner equation generalized for any gradient shape—an overview of most pulse sequences measuring free diffusion. *Concepts Magn. Reson. Part A* **40A**, 39–65 (2012).

91. Roosen-Runge, F. *et al.* Protein diffusion in crowded electrolyte solutions. *Biochim. Biophys. Acta - Proteins Proteomics* **1804**, 68–75 (2010).

92. Roosen-Runge, F. *et al.* Protein self-diffusion in crowded solutions. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 11815–11820 (2011).

93. Grimaldo, M. *et al.* Protein Short-Time Diffusion in a Naturally Crowded Environment. *J. Phys. Chem. Lett.* **10**, 1709–1715 (2019).

94. Singh, P. S. Small-Angle Scattering Techniques (SAXS/SANS). *Membr. Charact.* 95–111 (2017) doi:10.1016/B978-0-444-63776-5.00006-1.

95. Stracy, M. *et al.* Transient non-specific DNA binding dominates the target search of bacterial DNA-binding proteins. *Mol. Cell* **81**, 1499-1514.e6 (2021).

96. Konopka, M. C. *et al.* Cytoplasmic protein mobility in osmotically stressed Escherichia coli. *J. Bacteriol.* **91**, 231–237 (2009).

97. Daddysman, M. K. & Fecko, C. J. Revisiting point FRAP to quantitatively characterize anomalous diffusion in live cells. *J. Phys. Chem. B* **117**, 1241–1251 (2013).

98. Weiss, M., Elsner, M., Kartberg, F. & Nilsson, T. Anomalous Subdiffusion Is a Measure for Cytoplasmic Crowding in Living Cells. *Biophys. J.* **87**, 3518 (2004).

99. Banks, D. S., Tressler, C., Peters, R. D., Höfling, F. & Fradin, C. Characterizing anomalous diffusion in crowded polymer solutions and gels over five decades in time with variable-lengthscale fluorescence correlation spectroscopy. *Soft Matter* **12**, 4190–4203 (2016).

100. Hou, S., Exell, J. & Welsher, K. Real-time 3D single molecule tracking. *Nat. Commun.*

2020 111 **11**, 1–10 (2020).

101. Gwosch, K. C. *et al.* MINFLUX nanoscopy delivers 3D multicolor nanometer resolution in cells. *Nat. Methods* **17**, 217–224 (2020).

102. Zipfel, W. R. *et al.* Live tissue intrinsic emission microscopy using multiphoton-excited native fluorescence and second harmonic generation. *Proc. Natl. Acad. Sci.* **100**, 7075–7080 (2003).

103. Cheung, M. S., Klimov, D. & Thirumalai, D. Molecular crowding enhances native state stability and refolding rates of globular proteins. *Proc. Natl. Acad. Sci.* **102**, 4753–4758 (2005).

104. Minh, D. D. L., Chang, C. E., Trylska, J., Tozzini, V. & McCammon, J. A. The Influence of Macromolecular Crowding on HIV-1 Protease Internal Dynamics. *J. Am. Chem. Soc.* **128**, 6006–6007 (2006).

105. Oh, I., Choi, S., Jung, Y. & Kim, J. S. Unusual size-dependence of effective interactions between collapsed polymers in crowded environments. *Soft Matter* **10**, 9098–9104 (2014).

106. Elcock, A. H. Atomic-level observation of macromolecular crowding effects: Escape of a protein from the GroEL cage. *Proc. Natl. Acad. Sci.* **100**, 2340–2344 (2003).

107. Kurniawan, N. A., Enemark, S. & Rajagopalan, R. Crowding alters the folding kinetics of a β-hairpin by modulating the stability of intermediates. *J. Am. Chem. Soc.* **134**, 10200–10208 (2012).

108. Trovato, F. & Tozzini, V. Diffusion within the cytoplasm: A mesoscale model of interacting macromolecules. *Biophys. J.* **107**, 2579–2591 (2014).

109. Obrien, E. P., Straub, J. E., Brooks, B. R. & Thirumalai, D. Influence of nanoparticle size and shape on oligomer formation of an amyloidogenic peptide. *J. Phys. Chem. Lett.* **2**, 1171–1177 (2011).

110. Bille, A., Linse, B., Mohanty, S. & Irbäck, A. Equilibrium simulation of trp-cage in the presence of protein crowders. *J. Chem. Phys.* **143**, 175102 (2015).

111. Mereghetti, P. & Wade, R. C. Brownian dynamics simulation of protein diffusion in crowded environments. *AIP Conf. Proc.* **1518**, 511–516 (2013).

112. Mereghetti, P. & Wade, R. C. Atomic detail brownian dynamics simulations of concentrated protein solutions with a mean field treatment of hydrodynamic interactions. *J. Phys. Chem. B* **116**, 8523–8533 (2012).

113. von Bülow, S., Siggel, M., Linke, M. & Hummer, G. Dynamic cluster formation determines viscosity and diffusion in dense protein solutions. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 9843–9852 (2019).

114. Yu, I. *et al.* Biomolecular interactions modulate macromolecular structure and dynamics in atomistic model of a bacterial cytoplasm. *Elife* **5**, (2016).

115. Nawrocki, G., Wang, P. H., Yu, I., Sugita, Y. & Feig, M. Slow-Down in Diffusion in Crowded Protein Solutions Correlates with Transient Cluster Formation. *J. Phys. Chem. B* **121**, 11072–11084 (2017).

116. Mereghetti, P., Gabdoulline, R. R. & Wade, R. C. Brownian dynamics simulation of protein solutions: Structural and dynamical properties. *Biophys. J.* **99**, 3782–3791 (2010).

117. Bicout, D. J. & Field, M. J. Stochastic dynamics simulations of macromolecular diffusion in a model of the cytoplasm of Escherichia coli. *J. Phys. Chem.* **100**, 2489–2497 (1996).

118. Goodsell, D. S. Inside a living cell. *Trends Biochem. Sci.* **16**, 203–206 (1991).

119. Rosenberg, R. O., Thirumalai, D. & Mountain, R. D. Liquid, crystalline and glassy states of binary charged colloidal suspensions. *J. Phys. Condens. Matter* **1**, 2109 (1989).

120. Gabdoulline, R. R. & Wade, R. C. Effective Charges for Macromolecules in Solvent. *J. Phys. Chem.* **100**, 3868–3878 (1996).

121. Hasnain, S., McClendon, C. L., Hsu, M. T., Jacobson, M. P. & Bandyopadhyay, P. A new coarse-grained model for E. coli cytoplasm: Accurate calculation of the diffusion coefficient of proteins and observation of anomalous diffusion. *PLoS One* **9**, e106466

(2014).

122. Feig, M. *et al.* Complete atomistic model of a bacterial cytoplasm for integrating physics, biochemistry, and systems biology. *J. Mol. Graph. Model.* **58**, 1–9 (2015).

123. Fiser, A., Do, R. K. G. & Šali, A. Modeling of loops in protein structures. *Protein Sci.* **9**, 1753–1773 (2000).

124. Šali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815 (1993).

125. Jung, J. *et al.* GENESIS: a hybrid-parallel and multi-scale molecular dynamics simulator with enhanced sampling algorithms for biomolecular and cellular simulations. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **5**, 310–323 (2015).

126. Best, R. B. *et al.* Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone φ, ψ and side-chain χ(1) and χ(2) dihedral angles. *J. Chem. Theory Comput.* **8**, 3257–3273 (2012).

127. Vanommeslaeghe, K. *et al.* CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **31**, 671–690 (2010).

128. Elcock, A. H. Models of macromolecular crowding effects and the need for quantitative comparisons with experiment. *Curr. Opin. Struct. Biol.* **20**, 196–206 (2010).

129. Ho, B., Baryshnikova, A. & Brown, G. W. Unification of Protein Abundance Datasets Yields a Quantitative Saccharomyces cerevisiae Proteome. *Cell Syst.* **6**, 192-205.e3 (2018).

130. Brancolini, G., Lopez, H., Corni, S. & Tozzini, V. Low-Resolution Models for the Interaction Dynamics of Coated Gold Nanoparticles with β2-microglobulin. *Int. J. Mol. Sci. 2019, Vol. 20, Page 3866* **20**, 3866 (2019).

131. Weeks, E. R. & Weitz, D. A. Subdiffusion and the cage effect studied near the colloidal glass transition. *Chem. Phys.* **284**, 361–367 (2002).

132. Doliwa, B. & Heuer, A. Cage effect, local anisotropies, and dynamic heterogeneities at the glass transition: A computer study of hard spheres. *Phys. Rev. Lett.* **80**, 4915–4918 (1998).

133. Yu, X. *et al.* WebSDA: A web server to simulate macromolecular diffusional association. *Nucleic Acids Res.* **43**, W220–W224 (2015).

134. Wang, J., Cieplak, P. & Kollman, P. A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.* **21**, 1049 (2000).

135. Jurrus, E. *et al.* Improvements to the APBS biomolecular solvation software suite. *Protein Sci.* **27**, 112–128 (2018).

136. Allen, M. P. & Tildesley, D. J. *Computer Simulation of Liquids.* (Oxford University Press, 2017).

137. Goiko, M., De Bruyn, J. R. & Heit, B. Short-Lived Cages Restrict Protein Diffusion in the Plasma Membrane. *Sci. Rep.* **6**, 1–13 (2016).

138. Xue, C., Zheng, X., Chen, K., Tian, Y. & Hu, G. Probing Non-Gaussianity in Confined Diffusion of Nanoparticles. *J. Phys. Chem. Lett.* **7**, 514–519 (2016).

139. Porcar, L. *et al.* Formation of the dynamic clusters in concentrated Lysozyme protein solutions. *J. Phys. Chem. Lett.* **1**, 126–129 (2010).

140. Etoc, F. *et al.* Non-specific interactions govern cytosolic diffusion of nanosized objects in mammalian cells. *Nat. Mater.* **17**, 740–746 (2018).

141. Lawley, S. D. Extreme statistics of anomalous subdiffusion following a fractional Fokker-Planck equation: Subdiffusion is faster than normal diffusion. *J. Phys. A Math. Theor.* **53**, 32 (2020).

142. Haugh, J. M. Analysis of reaction-diffusion systems with anomalous subdiffusion. *Biophys. J.* **97**, 435–442 (2009).

143. Van Den Berg, J., Boersma, A. J. & Poolman, B. Microorganisms maintain crowding homeostasis. *Nat. Rev. Microbiol.* **15**, 309–318 (2017).

144. Waldron, C. & Lacroute, F. Effect of growth rate on the amounts of ribosomal and transfer ribonucleic acids in yeast. *J. Bacteriol.* **122**, 855–865 (1975).

145. Yamaguchi, M. *et al.* Structome of Saccharomyces cerevisiae determined by freeze-substitution and serial ultrathin-sectioning electron microscopy. *J. Electron Microsc. (Tokyo).* **60**, 321–335 (2011).

146. Wei, D. *et al.* High-resolution three-dimensional reconstruction of a whole yeast cell using focused-ion beam scanning electron microscopy. *Biotechniques* **53**, 41–48 (2012).

147. Berriz, G. F., Beaver, J. E., Cenik, C., Tasan, M. & Roth, F. P. Next generation software for functional trend analysis. *Bioinformatics* **25**, 3043–3044 (2009).

148. Sasidharan, K., Amariei, C., Tomita, M. & Murray, D. B. Rapid DNA, RNA and protein extraction protocols optimized for slow continuously growing yeast cultures. *Yeast* **29**, 311–322 (2012).

149. Yofe, I. *et al.* One library to make them all: Streamlining the creation of yeast libraries via a SWAp-Tag strategy. *Nat. Methods* **13**, 371–378 (2016).

150. Lu, P., Vogel, C., Wang, R., Yao, X. & Marcotte, E. M. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.* **25**, 117–124 (2007).

151. Peng, M. *et al.* Protease bias in absolute protein quantitation. *Nature Methods* vol. 9 524–525 (2012).

152. Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N. & Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **11**, 319–324 (2014).

153. Lawless, C. *et al.* Direct and absolute quantification of over 1800 yeast proteins via selected reaction monitoring. *Mol. Cell. Proteomics* **15**, 1309–1322 (2016).

154. Lahtvee, P. J. *et al.* Absolute Quantification of Protein and mRNA Abundances Demonstrate Variability in Gene-Specific Translation Efficiency in Yeast. *Cell Syst.* **4**, 495-504.e5 (2017).

155. De Godoy, L. M. F. *et al.* Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **455**, 1251–1254 (2008).

156. Picotti, P. *et al.* A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. *Nature* **494**, 266–270 (2013).

157. Lee, M. V. *et al.* A dynamic model of proteome changes reveals new roles for transcript alteration in yeast. *Mol. Syst. Biol.* **7**, (2011).

158. Thakur, S. S. *et al.* Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation. *Mol. Cell. Proteomics* **10**, M110.003699 (2011).

159. Nagaraj, N. *et al.* System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top orbitrap. *Mol. Cell. Proteomics* **11**, M111.013722 (2012).

160. Webb, K. J., Xu, T., Park, S. K. & Yates, J. R. Modified MuDPIT separation identified 4488 proteins in a system-wide analysis of quiescence in yeast. *J. Proteome Res.* **12**, 2177–2184 (2013).

161. Tkach, J. M. *et al.* Dissecting DNA damage response pathways by analysing protein localization and abundance changes during DNA replication stress. *Nat. Cell Biol.* **14**, 966–976 (2012).

162. Breker, M., Gymrek, M. & Schuldiner, M. A novel single-cell screening platform reveals proteome plasticity during yeast stress responses. *J. Cell Biol.* **200**, 839–850 (2013).

163. Dénervaud, N. *et al.* A chemostat array enables the spatio-temporal analysis of the yeast proteome. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 15842–15847 (2013).

164. Mazumder, A., Pesudo, L. Q., McRee, S., Bathe, M. & Samson, L. D. Genome-wide single-cell-level screen for protein abundance and localization changes in response to DNA damage in S. cerevisiae. *Nucleic Acids Res.* **41**, 9310–9324 (2013).

165. Chong, Y. T. *et al.* Yeast proteome dynamics from single cell imaging and automated analysis. *Cell* **161**, 1413–1424 (2015).

166. Newman, J. R. S. *et al.* Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature* **441**, 840–846 (2006).

167. Lee, M. W. *et al.* Global protein expression profiling of budding yeast in response to DNA damage. *Yeast* **24**, 145–154 (2007).

168. Davidson, G. S. *et al.* The proteomics of quiescent and nonquiescent cell differentiation in yeast stationary-phase cultures. *Mol. Biol. Cell* **22**, 988–998 (2011).

169. Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425**, 737–741 (2003).

170. Perry, R. P. Balanced production of ribosomal proteins. *Gene* vol. 401 1–3 (2007).

171. Warner, J. R. The economics of ribosome biosynthesis in yeast. *Trends Biochem. Sci.* **24**, 437–440 (1999).

172. Ben-Shem, A. *et al.* The structure of the eukaryotic ribosome at 3.0 Å resolution. *Science (80-.).* **334**, 1524–1529 (2011).

173. Lu, H., Zhu, Y. F., Xiong, J., Wang, R. & Jia, Z. Potential extra-ribosomal functions of ribosomal proteins in Saccharomyces cerevisiae. *Microbiol. Res.* **177**, 28–33 (2015).

174. von der Haar, T. A quantitative estimation of the global translational activity in logarithmically growing yeast cells. *BMC Syst. Biol.* **2**, (2008).

175. Jorgensen, P., Nishikawa, J. L., Breitkreutz, B. J. & Tyers, M. Systematic identification of pathways that couple cell growth and division in yeast. *Science (80-.).* **297**, 395–400 (2002).

176. Ellis, R. J. Macromolecular crowding: an important but neglected aspect of the intracellular environment. *Curr. Opin. Struct. Biol.* **11**, 114–119 (2001).

177. Miermont, A. *et al.* Severe osmotic compression triggers a slowdown of intracellular signaling, which can be explained by molecular crowding. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 5725–5730 (2013).

178. Hohmann, S. Control of high osmolarity signalling in the yeast Saccharomyces cerevisiae. *FEBS Lett.* **583**, 4025–4029 (2009).

179. Konopka, M. C. *et al.* Cytoplasmic protein mobility in osmotically stressed Escherichia coli. *J. Bacteriol.* **91**, 231–237 (2009).

180. Guex, N. & Peitsch, M. C. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **18**, 2714–2723 (1997).

181. BIOVIA & Dassault Systèmes. Discovery Studio Visualizer,. (2020).

182. Huang, J. & Mackerell, A. D. CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *J. Comput. Chem.* **34**, 2135–2145 (2013).

183. Xu, Y., Vanommeslaeghe, K., Aleksandrov, A., MacKerell, A. D. & Nilsson, L. Additive CHARMM force field for naturally occurring modified ribonucleotides. *J. Comput. Chem.* **37**, 896–912 (2016).

184. Dolinsky, T. J., Nielsen, J. E., McCammon, J. A. & Baker, N. A. PDB2PQR: An automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.* **32**, W655-7 (2004).

185. Dolinsky, T. J. *et al.* PDB2PQR: Expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.* **35**, W522–W525 (2007).

186. Fernández-Niño, M. *et al.* The cytosolic pH of individual Saccharomyces cerevisiae cells is a key factor in acetic acid tolerance. *Appl. Environ. Microbiol.* **81**, 7813–7821 (2015).

187. Holst, M. & Saied, F. Multigrid solution of the Poisson-Boltzmann equation. *J. Comput. Chem.* **14**, 105–113 (1993).

188. Bank, R. E. & Holst, M. A New Paradigm for Parallel Adaptive Meshing Algorithms. *http://dx.doi.org/10.1137/S003614450342061* **45**, 291–323 (2006).

189. Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc. Natl. Acad. Sci.* **98**, 10037–10041 (2001).

190. Cohen, A. E. & Venkatachalam, V. Bringing Bioelectricity to Light. (2014) doi:10.1146/annurev-biophys-051013-022717.

191. Nissen, P. *et al.* Crystal structure of the ternary complex of Phe-tRNAPhe, EF-Tu, and a GTP analog. *Science (80-.).* **270**, 1464–1472 (1995).

192. Waterhouse, A. *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018).

193. Bienert, S. *et al.* The SWISS-MODEL Repository-new features and functionality. *Nucleic Acids Res.* **45**, D313–D319 (2017).

194. Dreher, T. W., Uhlenbeck, O. C. & Browning, K. S. Quantitative Assessment of EF-1α·GTP Binding to Aminoacyl-tRNAs, Aminoacyl-viral RNA, and tRNA Shows Close Correspondence to the RNA Binding Properties of EF-Tu *. *J. Biol. Chem.* **274**, 666–672 (1999).

195. Zhang, Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* **9**, 1–8 (2008).

196. Roy, A., Kucukural, A. & Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc. 2010 54* **5**, 725–738 (2010).

197. Yang, J. *et al.* The I-TASSER Suite: protein structure and function prediction. *Nat. Methods* **12**, 7–8 (2015).

198. Xu, D. & Zhang, Y. Toward optimal fragment generations for ab initio protein structure assembly. *Proteins* **81**, 229 (2013).

199. Xu, D. & Zhang, Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins* **80**, 1715–1735 (2012).

200. Evans, M. E., Clark, W. C., Zheng, G. & Pan, T. Determination of tRNA aminoacylation levels by high-throughput sequencing. *Nucleic Acids Res.* **45**, e133 (2017).

201. Gromadski, K. B. *et al.* Kinetics of the interactions between yeast elongation factors 1A and 1Bα, guanine nucleotides, and aminoacyl-tRNA. *J. Biol. Chem.* **282**, 35629–

35637 (2007).

202. Liarzi, O. & Epel, B. L. Development of a quantitative tool for measuring changes in the coefficient of conductivity of plasmodesmata induced by developmental, biotic, and abiotic signals. *Protoplasma* **225**, 67–76 (2005).

203. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583 (2021).

204. Skolnick, J., Gao, M., Zhou, H. & Singh, S. AlphaFold 2: Why It Works and Its Implications for Understanding the Relationships of Protein Sequence, Structure, and Function. *J. Chem. Inf. Model.* **61**, 4827–4831 (2021).

205. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).

206. Robbins, E., Pederson, T. & Klein, P. Comparison of mitotic phenomena and effects induced by hypertonic solutions in hela cells. *J. Cell Biol.* **44**, 400–416 (1970).

207. Fluitt, A., Pienaar, E. & Viljoen, H. Ribosome kinetics and aa-tRNA competition determine rate and fidelity of peptide synthesis. *Comput. Biol. Chem.* **31**, 335–346 (2007).

208. Park, S. J., Ahn, H. S., Kim, J. S. & Lee, C. Evaluation of multi-tRNA synthetase complex by multiple reaction monitoring mass spectrometry coupled with size exclusion chromatography. *PLoS One* **10**, e0142253 (2015).

209. Harris, C. L. & Kolanko, C. J. Aminoacyl-tRNA synthetase complex in Saccharomyces cerevisiae. *Biochem. J.* **309**, 321–324 (1995).

210. David, A. *et al.* RNA binding targets aminoacyl-tRNA synthetases to translating ribosomes. *J. Biol. Chem.* **286**, 20688–20700 (2011).

211. Iborra, F. J., Pombo, A., Jackson, D. A. & Cook, P. R. Active RNA polymerases are localized within discrete transcription ‘factories’ in human nuclei. *J. Cell Sci.* **109**, 1427–1436 (1996).

Every reasonable effort has been made to acknowledge the owners of copyright material. I would be pleased to hear from any copyright owner who has been omitted or incorrectly acknowledged.

Appendix

Content permission

The contents of the chapter 4 were reproduced under the terms of the Creative Commons Attribution License (CC-BY 4.0). The title of the paper, from which the contents were taken, is “Definition of the Minimal Contents for the Molecular Simulation of the Yeast Cytoplasm” (Kompella, V. P. S., Stansfield, I., Romano, M. C., & Mancera, R. L. (2019). Frontiers in Molecular Biosciences, 6, 97. https://doi.org/10.3389/fmolb.2019.00097).

Figure permissions

1. Reprinted from Biophysical Journal, Vol 94, Ridgway, D., Broderick, G., Lopez-Campistrous, A., Ru’aini, M., Winter, P., Hamilton, M., Boulanger, P., Kovalenko, A., Ellison, M.J., Coarse-Grained Molecular Simulation of Diffusion and Reaction Kinetics in a Crowded Virtual Cytoplasm, 3748–3759., Copyright (2008), with permission from Elsevier [OR APPLICABLE SOCIETY COPYRIGHT OWNER].

2. Reprinted from Biophysical Journal, Vol 107, Trovato, F., Tozzini, V., Diffusion within the cytoplasm: A mesoscale model of interacting macromolecules, 2579–2591., Copyright (2014), with permission from Elsevier [OR APPLICABLE SOCIETY COPYRIGHT OWNER]."

ELSEVIER LICENSE
TERMS AND CONDITIONS
Jan 29, 2022

This Agreement between Mr. Vijay Phanindra Srikanth Kompella ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	5212310580285
License date	Dec 19, 2021
Licensed Content Publisher	Elsevier

Licensed Content Publication	Biophysical Journal
Licensed Content Title	Coarse-Grained Molecular Simulation of Diffusion and Reaction Kinetics in a Crowded Virtual Cytoplasm
Licensed Content Author	Douglas Ridgway,Gordon Broderick,Ana Lopez-Campistrous,Melania Ru'aini,Philip Winter,Matthew Hamilton,Pierre Boulanger,Andriy Kovalenko,Michael J. Ellison
Licensed Content Date	May 15, 2008
Licensed Content Volume	94
Licensed Content Issue	10
Licensed Content Pages	12
Start Page	3748
End Page	3759
Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Title	PhD student
Institution name	Curtin University
Expected presentation date	Jan 2022
Order reference number	1
Portions	Figure 2
Requestor Location	Mr. Vijay Phanindra Srikanth Kompella 79 Kilkenny circle Waterford Perth, WA 6152 Australia Attn: Mr. Vijay Phanindra Srikanth Kompella
Publisher Tax ID	GB 494 6272 12
Total	0.00 AUD

ELSEVIER LICENSE
TERMS AND CONDITIONS
Jan 29, 2022

This Agreement between Mr. Vijay Phanindra Srikanth Kompella ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	5212330556264
License date	Dec 19, 2021
Licensed Content Publisher	Elsevier
Licensed Content Publication	Biophysical Journal
Licensed Content Title	Diffusion within the Cytoplasm: A Mesoscale Model of Interacting Macromolecules
Licensed Content Author	Fabio Trovato,Valentina Tozzini
Licensed Content Date	Dec 2, 2014
Licensed Content Volume	107
Licensed Content Issue	11
Licensed Content Pages	13
Start Page	2579
End Page	2591
Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Title	PhD student
Institution name	Curtin University
Expected presentation date	Jan 2022
Order reference number	2
Portions	figure 6
Requestor Location	Mr. Vijay Phanindra Srikanth Kompella 79 Kilkenny circle Waterford Perth, WA 6152

	Australia Attn: Mr. Vijay Phanindra Srikanth Kompella
Publisher Tax ID	GB 494 6272 12
Total	0.00 AUD