Resolving the Protein Folding Problem Using Clustering Partition and Random Walk Theory with Extended Applications in Alzheimer's Research using Galton Watson Processes and Stochastic Theory

Dwaipayan Basu[1*], Suvadip Sana[2]

[1]Department of Microbiology, St. Xavier's College (Autonomous), Kolkata, India

[2]Department of Mathematics, Indian Statistical Institute, Bangalore, Karnataka 560059, India

ABSTRACT: Exploring the nature of protein folding, the following work is an elucidation of a bio-mathematical explanation of the Levinthal's paradox using Clustering Partition among different torsional states in conjunction with Random Walk Theory. The kinetics of protein folding mechanisms have been resolved by integrating Overlapping Clusters across different Hierarchical Levels of protein folding, from simplest (Primary) to the most folded (Tertiary) structural conformations by introducing stochastic perturbations. The analytical dynamics of attaining stable conformations and at the same time spontaneous loss of native structures pave the way to protein misfolding and the self-propagating phenomenon of toxic protein intermediates in neurodegenerative diseases. The pathophysiological manifestations of one such protein Amyloid-beta, implicated in Alzheimer's disease-meandering through states of the beta-amyloid fibril accumulation, has been addressed; along with an investigation into the tau protein's instability and related clinical conditions upon exposure to metal ion induced neurotoxicity. The two major hypotheses of Alzheimer's Disease: the 'Amyloid Cascade Theory' followed by the 'Microbial Infection Theory' have been scrutinized and unified on the basis of their convergence to genetic dysbiosis; associating unusual interactions among proteins as well as head towards the existence of a probable link between the protein modifications and microbe induced protein misfolds. This has been followed by an investigation into the dynamics of Amyloid fibril propagation, in a model analogous to Prion replication. The immense coverage of this bio-mathematical study can help decipher the complex molecular mechanisms underlying the process of protein folding and help supplement clinical investigations into the causes and remedies of Alzheimer's disease.

KEYWORDS: Levinthal's Paradox; Clustering Partition; Random Walk Theory; Stochastic Perturbations; Galton-Watson Process; Amyloid-beta; Alzheimer's Disease

https://doi.org/10.29294/IJASE.6.S1.2019.30-33

1. INTRODUCTION

An extension of the 'Protein-Folding Problem'- thefamed 'Levinthal's Paradox'; named after the American molecular biologist who first proposed the problem, is one of the most archaic yet unresolved mysteries of the Biological world [1]. The kinetics of protein folding were studied intently throughout the late 20th century, yet a concrete resolution of the paradox was not proposed until the work of Edward Trifonov who suggested a possible explanation using the concept of autonomously aggregating protein 'domains' serving as 'Nucleation Points'- but his approach did not encompass kinetic studies, and was thus incomplete in scope and application [2]. The significance of the problem comes into focus because the mechanism and molecular dynamics of protein folding are yet to be deciphered completely; moreover, the pathophysiological implications of misfolded and self-aggregating proteins such as the Amyloid-Beta (Aβ) implicated in Alzheimer's Disease are serious and often irreversible. We have provided an accurate and detailed mathematical model resolving the Levinthal's Paradox using Clustering Partition Theory and Random Walk Theory. Moreover, we have provided a statistical analysis of heavy metal induced neurotoxicity on the aggregation properties of Aβ and finally provided a connecting link between the two leading hypotheses in Alzheimer's research: the 'Amyloid Cascade Hypothesis' and the 'Infection Hypothesis'. This has led us to a unified theory explaining the progression and incidence of Alzheimer's on the principle of 'Genetic Dysbiosis'…. which we have explored from the perspective of Theoretical Biology i.e. Biomathematics in conjunction with a detailed biochemical and molecular biological model.

Alzheimer's, one of the leading causes of death among elderly people, is irreversible and incurable. Its onset is also difficult to detect, owing to lack of apparent symptoms and thus is nearly impossible to control once pathophysiological manifestations finally present themselves. By statistical studies, we can predict the probability and genetic triggers that can lead to onset of Alzheimer's over time- which might prove to be an invaluable tool in research into this condition.

The aim of the following work is to explore and provide a suitable mathematical model to resolve one

**Corresponding Author: dwaipayanbasu27@gmail.com*

Received: 15.05.2019 **Accepted:** 18.06.2019 **Published on**: 20.07.2019

of the most notable extensions of the 'Protein-Folding Problem', i.e. the 'Levinthal's Paradox'. We also aim to decipher the reasons for misfolding and mis-aggregation of certain neurotoxic protein isoforms in accordance with the aforementioned model, with emphasis on Amyloid Beta(Aβ), implicated in Alzheimer's Disease. We follow this by providing a detailed Markov Model to explore the progression of Aβ plaque deposition with respect to probability and kinetics.

2. MATHEMATICAL FORMULATIONS

Clustering Partition

We develop the idea of 'Clustering Partition' to provide a plausible explanation to resolve the Levinthal's paradox. The idea is motivated by the theory of equivalence classes in abstract mathematics. Briefly, the Levinthal's paradox concerns protein folding kinetics, i.e. the theoretical time that should be required for a hypothetical polypeptide to assume it's 'native' conformation- contrasts sharply with the actual time required in vivo for proteins to fold (microseconds). [3]

We have used the same Hypothetical Polypeptide stated by Levinthal himself in the statement of his paradox of 100 amino acid residues. This protein would thus contain 99 peptide bonds, with 198 phi and psi angles-each of which being a sigma bond, could freely rotate and based on torsional constraints, choose among 3 possible conformations. Thus, the total number of possible conformations to choose from is 3^{198}.[4]

- We denote S as our state space.
- Thus, the number of elements of S is 3^{198}.
  Claim: Now we may talk of 'linkages.'
- In combinatorics, linkage is the association of one permutational cluster with another.
- Now, it is a valid assumption from Modularity, that the protein will assume cluster in a hierarchical manner.
  Say, initially we partition our state space into three disjoint partition sets, each containing equal number of states. and denote them as:
- S_{11}, S_{12} and S_{13}.
  So mathematically $S = \sum_{i=1}^{3} S_{1i}$
  Here 1 denotes the first selection process. The hypothetical favored state is in one of these three sets.
- Assume it is in S_{11}.
  Also note: Each of these partition sets have 3^{197} state spaces each.
  Now note: number of states has been reduced from 3^{198} to 3^{197}.
  Now we again break the space S_{11} into three disjoint sets.
  Following the same algorithm, we get:
  Three disjoint sets
- S_{21}, S_{22}, S_{23}.
  Here 2 denotes the second selection process.
  Also note: Each of these sets contain 3^{196} state spaces each.

Generalizing these results, we obtain:

- If available conformers= X^Y
- Then total samplings required= (Y-1) +X
- So, for the previous example i.e. 3^{198}
- Total Samplings= (198-1) +3=200
- Time required= 200
- $200*10^{-9}$s= 0.2 microseconds

This result, i.e. 0.2 microseconds, is consistent with the experimental folding duration observed for most mammalian proteins in vivo.

However, we have not illustrated the basis of selection for each cluster at a particular hierarchical level. This can be tackled using a result from probability theory termed 'Random Walk Theory'.

Random Walk Theory

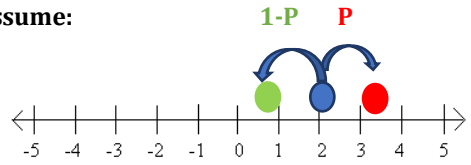
For this purpose, we need to define the terms 'recurrent' and 'transient' on the basis of probability. [5]

Transient: A state which is visited with probability 1 by a finite number of Random Walks.

Recurrent: A state which is visited with probability 1 by an infinite number of random walks.

For reasons of simplicity, we have limited our discussion to one dimensional Random Walks, though it is equally applicable in N-Dimensions.

Assume:



Thus, if each digit be thought to represent a conformational state in a particular hierarchical level, then a partially ordered protein molecule in any one of these states say 0, will have 2 degrees of freedom; i.e. it can migrate to the states -1 and 1. Since these two events are exhaustive, the sum of their probabilities will be equal to 1.

Claim: One of the states in this series is 'recurrent' while all others are 'transient'.

(N.B It is a valid presupposition that P≠1/2. If P were equal to ½, then all states would be transient, and the protein would not fold spontaneously).

We denote, each of the Hierarchical Levels as: H_0, H_1, H_2…… H_N

Now Each Level contains a finite number of conformations denoted as a_{ij} where i represent the Hierarchical Level and j the position in each.

Thus, we obtain:

H_0: $a_{0,0}$; $a_{0,1}$; $a_{0,3}$……

H_1: $a_{1,0}$; $a_{1,1}$; $a_{1,3}$……

H_2: $a_{2,0}$; $a_{2,1}$; $a_{2,3}$……

.

.

.

H_N: $a_{N,0}$; $a_{N,1}$; $a_{N,3}$……

So, here in each level of aggregation, one state is selected on the basis of Random Walk Theory, and this is repeated throughout all of the hierarchical levels, thus drastically reducing the effective number of samplings necessary to arrive at the favored conformation- increasing the kinetic efficiency of protein folding.
(N.B. We have not included the chaperone-assisted protein folding models in this investigation for the sake of simplicity. However, it is hypothesized that this will follow similar dynamics: as chaperones simply act to facilitate thermodynamics-mediated lowering of entropy, and thus do not seem to play a significant role in altering the kinetics of protein folding).

Implications in Alzheimer's Research

According to the Amyloid Cascade Hypothesis of Alzheimer's Disease, the protein Amyloid Beta (Aβ) accumulation triggers neural decay and formation of Neurofibrillary Tangles (NFT's), a common manifestation of the disease. The method of Amyloidogenesis in Alzheimer's can be compared to the propagation of the prion protein[6]which is also implicated in a number of fatal neurodegenerative diseases, collectively referred to as the TSE's (Transmissible Spongiform Encephalopathies).

It is clearly evident from our previous arguments regarding 'Transient" and 'Recurrent" states in protein ensembles-that the Aβ oligomers get conformationally entrapped in a torsional state other than the Recurrent State and thus induces misfolding (By a mechanism we hypothesize to be similar to 'steric chaperones') of other Aβ oligomers.

This 'seeding' mechanism, whereby a single neurotoxic, non-fibrillar Aβ oligomer induces conformational instability in a non-toxic partner, can be elucidated using a special Markov Process, termed "Galton-Watson Process".

3. MATHEMATICAL FORMULATIONS

Markov Process

Let x_1, x_2, x_3,……x_n be a sequence of random variablesindexed by natural numbersassuming values in a set S which is a countable set.
So, we can enumerate S as:
$S= \{i_1, i_2, i_3,……\}$
We say that the sequence is a Markov process if future is independent of past given present.[5] We can rephrase it in terms of 'Condition probability' as follows:
$P [X_{n+1}=j_{n+1} / X_n=J_n, X_{n-1}=J_{n-1},…., X_1=J_1]= P [X_{n+1}=J_{n=1} / X_n=J_n]$

Model Making

We will see how a single Aβ oligomer can give rise to another set of toxic oligomeric intermediates. We will mostly be interested in asymptotic behaviour.
Let us assume that in the 0th generation we have only 1 Aβ oligomer. No of progenies it can have follows a Poisson Distribution with a mean λ.
So, let X_1= Number of progenies of 1 Aβ oligomer
now, let X_2= Number of progenies of the 1st generation of oligomers.
……..
let X_n= Number of progenies of the (n-1)th generation of oligomers.
Then by our construction note that,
$X_{n+1}= Z_1+ Z_2+……+ Z_n= \sum_{i=1}^{Xn} Xi$
Here each $Z_i \sim$ poisson (λ)
Then $E[X_{n+1}]= \lambda E[X_n]$……{Hence E[Y] denotes the mean value of the random variable}
The above equation follows from a well-known theorem of probability.
Hence, $E[X_n]= \lambda^n$……{follows from Inductive argument}
So, you can see that if λ<1 then :
$\lim_{n\to\infty} \lambda \hat{} n = \lim_{n\to\infty} E[Xn] =0$
and if λ>1, then:
$\lim_{n\to\infty} E[Xn] =\infty$

So we can see that if λ<1, then the Expected value of X_n's converge to zero, and since X_n's are all non-negative random variables, after some generations the population stabilizes and the Amyloidogenesis cannot progress.

Moreover, we see that if λ>1, then the population explodes, there being a high chance of amyloid plaque accumulation-followed by neural decay and cell death.

Now the value of λ will most certainly be multifactorial, which lends credence to the theory of 'Genetic Dysbiosis' over either of the two previously established theories: namely the 'Amyloid Cascade Hypothesis' as well the 'Infection Hypothesis'- leading us to conclude that the onset of Alzheimer's is most possibly triggered by a wide variety of genetic and environmental factors acting in conjunction.

4. CONCLUSION

Thereby it is possible to predict, with a reasonable degree of accuracy, the probability and kinetics of plaque deposition in neural tissue in Alzheimer 's disease, once we have deciphered the value of λ. However, this presents additional challenges which shall not be scrutinized in this particular paper. It should suffice to say that it may be determined using advanced sampling theory. In this work, we have therefore explored the biomathematical model governing the molecular dynamics of protein folding, and we have also addressed the misfolding phenomenon of certain proteins in light of this model. We have also suggested a model for the propagation of one such misfolded protein variant, termed Aβ, implicated in Alzheimer 's disease.

REFERENCES

[1]. Zwanzig R, Szabo A, Bagchi B. Levinthal's paradox. *Proc Natl Acad Sci U S A*. 1992;89(1):20–22.
[2]. Igor N. Berezovsky & Edward N. Trifonov (2002) Loop Fold Structure of Proteins: Resolution of Levinthal's Paradox, Journal of Biomolecular Structure and Dynamics, 20:1, 5-6,

[3]. William A. Eaton, Victor Muñoz, Stephen J. Hagen, GauriS. , Jas Lisa J. Lapidus, Eric R. Henry, James Hofrichter 2000. Fast Kinetics and Mechanisms in Protein Folding. Annual Review of Biophysics and Biomolecular Structure, 29:1, 327-359

[4]. C. Levinthal, "How to Fold Graciously," Mossbauer Spectroscopy in Biological Systems Proceedings, 67(41), 1969, pp. 22-26.

[5]. Ross, S.M., 1983. Stochastic Processes. Wiley.

[6]. Basu, D., Sana, S., 2018. Emerging Trends in Science and Technology. Narosa, India, pp. 33.1-33.5

Selection and/or Peer-review under the responsibility of 2nd International Students' Conference on Innovations in Science and Technology (Spectrum – 2019), Kolkata