EUROPHYSICS LETTERS 15 November 2004

Europhys. Lett., **68** (4), pp. 589–595 (2004)
DOI: 10.1209/epl/i2004-10249-7

A model for the accidental catalysis of protein unfolding *in vivo*

R. P. Sear(*)
Department of Physics, University of Surrey - Guildford, Surrey GU2 7XH, UK

received 13 May 2004; accepted in final form 28 September 2004
published online 20 October 2004

PACS. 87.15.Nn – Properties of solutions; aggregation and crystallization of macromolecules.
PACS. 87.14.Ee – Proteins.
PACS. 87.15.He – Dynamics and conformational changes.

Abstract. – Activated processes such as protein unfolding are highly sensitive to heterogeneity in the environment. We study a highly simplified model of a protein in a random heterogeneous environment, a model of the *in vivo* environment. It is found that if the heterogeneity is sufficiently large, the total rate of the process is essentially a random variable; this may be the cause of the species-to-species variability in the rate of prion protein conversion found by Deleault *et al.* (*Nature*, **425** (2003) 717).

Protein unfolding is implicated in a number of diseases including prion diseases such as Creutzfeldt-Jakob disease [1–3]. It is an activated process, a free-energy barrier must be overcome for a protein to unfold from its native state. At the top of the barrier the protein is in the transition state for unfolding, and the transition state's free energy determines the rate [4, 5]. As the rate depends exponentially on the free energy, the rate is very sensitive to interactions of other molecules from the environment with the transition state. Inside living cells there is a mixture of thousands of different proteins, RNA, etc.; if any of them can interact with the transition state of unfolding such that its free energy is only a few $k_{\mathrm{B}}T$ lower, then the rate of prion protein conversion when interacting with this other molecule will be increased by an order of magnitude. Supattapone and coworkers [6] studied prion conversion in cell extracts and found that the rate of prion protein conversion was greatly accelerated by an RNA molecule or molecules, and that surprisingly this acceleration was specific to the RNA of only some species. Here we look at a very simple model of unfolding *in vivo*, and examine how the rate of unfolding is affected by the protein being in a complex mixture of many other molecules. Characterising the interactions of thousands of different molecules with the transition state is a hopeless task and so we resort to a statistical approach [7, 8]. We take the interactions to be random variables. This reduces the problem from characterising a huge number of interactions to just characterising the distribution function of these random variables. By taking all the interactions to be random variables we are ignoring the fact that natural selection may be acting to restrict or increase the strength of some of the interactions, and so our model will be a poor one if the RNA accelerating the rate of prion protein conversion has evolved to interact

(*) E-mail: r.sear@surrey.ac.uk

strongly with the prion protein. Very little is definitely known about the function of the prion protein [2, 3] and so we cannot rule out this possibility. We find that if the free energies of interaction with the transition state are spread over a wide range, unfolding occurs predominantly with the transition state in contact with one or a few of the other molecules present. These molecules are the ones responsible for the outliers of the distribution of interactions with the transition state, they are the ones that interact most strongly with the transition state. If we take these outliers to be RNA molecules, then the predictions of our model are consistent with the experimental findings of Supattapone and coworkers [6]. When one or a few outliers dominate the rate, it may vary significantly from species to species simply due to chance species-to-species variations in the nucleotide or amino-acid sequences of these outliers.

Supattapone and coworkers [6] have shown that the conversion of a prion protein from the PrP^C form to the PrPres form is greatly accelerated by a specific RNA molecule or by a small set of such molecules. The PrP^C form is the normal form while the PrPres form is analogous to the form associated with disease. The PrPres form is so-called because it is Protease RESistant, *i.e.*, not destroyed by the proteases that cut the chains of normal proteins. The two forms of the protein have the same amino-acid sequence, they differ only in conformation. The interconversion is known to be accelerated by PrPres itself but Supattapone and coworkers showed that a specific fraction of RNA molecules from both hamsters and mice but *not* the same fraction from invertebrates, also appeared to accelerate the conversion of the same protein. Of course, in terms of the prion diseases in different species the prion protein itself will vary from species to species and this will cause variability. Here we are considering variability not in the prion protein itself but in a cofactor that interacts with the prion protein. There is other experimental data on possible cofactors affecting the rate of prion protein conversion. Cordeiro *et al.* [9] suggest, on the basis of experimental evidence, that DNA reduces the free-energy barrier to the conversion of a prion protein into the form associated with the disease. Other work on prions has implicated as a cofactor not RNA but a protein-dubbed protein X [10]. There is considerable uncertainty surrounding the mechanism behind prion diseases [2]. See the reviews of Harris [2] and of Aguzzi and Polymenidou [3] for an introduction to prions.

We assume the unfolding of a protein to be a simple activated process [4, 5], its rate having an exponential dependence on the barrier to unfolding, ΔF^*: the difference in free energy between the folded protein and the transition state. The transition state being, by definition, the state of the protein along the unfolding pathway that has the highest free energy. Our model of the transition state for unfolding is a linear polymer on a simple cubic lattice, n_M monomers long. Inside a living cell, there are the surfaces of proteins, of membranes, of DNA etc. For simplicity, we lump all these surfaces together into a large flat surface which we model by a plane of lattice sites. A transition state in contact with a part of this surface is shown in fig. 1. The monomers of the transition state and of the surface are either hydrophilic or hydrophobic. We take B of the monomers of the transition state to be hydrophobic. We assume that unfolding proceeds by some part of the protein, n_M monomers long, unfolding, its free energy increasing as it does so until the free energy reaches a maximum at the transition state [11]. This transition state can contact the surface, as seen in fig. 1, and for each hydrophobic monomer of the transition state in contact with a hydrophobic monomer of the surface there is a contribution of $-\epsilon$ to the free energy of the transition state. The only energy of interaction is between hydrophobic monomers.

The surfaces are those of proteins, RNA, etc. and so are coded for by the genome of the organism. Thus they will differ between one species and another. We have no means of calculating them from the genome of an organism and so resort to modelling the surface with a purely random distribution of hydrophobic and hydrophilic monomers. Each monomer is hydrophobic with probability h and hydrophilic with probability $1 - h$. This is in the spirit

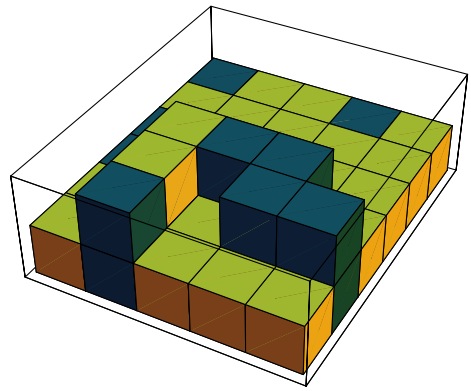

Fig. 1 – Schematic representation of our starting model for the transition state in contact with a patch of surface. The surface is assumed planar for simplicity. Hydrophobic monomers are shown as the dark cubes, and hydrophilic monomers are the light cubes. The transition state is the set of $n_M = 7$ contiguous monomers, $B = 5$ of which are hydrophobic, on top of the surface.

pioneered by Wigner and others [7] in random matrix theory, see ref. [8] for an application to protein mixtures. The surface provides N_s different positions and configurations of the transition state in which the transition state can interact with the surface, we call these unfolding configurations. We neglect any correlations between the interaction energy at different unfolding configurations on the surface and assume that the N_s configurations are independent. Then, if we denote the free energy of the transition state when it is not interacting with any other monomers by ΔF_0^*, the rate of unfolding at configuration i, R_i, is

$$R_i = \nu \exp\left[-\Delta F_0^* + n_i \epsilon\right], \tag{1}$$

where n_i is the number of hydrophobic monomers of the transition state that are adjacent to hydrophobic parts of the surface. Thus, the surfaces present are specified by the set of N_s values of the random variables n_i. Note that we have assumed that the attempt frequency ν is the same for all unfolding configurations, only the free-energy barrier varies. We use units such that the thermal energy $k_\mathrm{B}T = 1$.

The rate of unfolding averaged over all N_s possible configurations is

$$R = N_\mathrm{s}^{-1} \sum_{i=1}^{N_\mathrm{s}} R_i \,. \tag{2}$$

Although we have used the specific example of protein unfolding, quite generally the rates of activated process are given by equations with the form of eq. (1) and so our theory will apply quite generally to activated processes *in vivo*. Equations similar to eqs. (1) and (2) were employed by Karpov and Oxtoby [14] to study nucleation, an activated process like unfolding, in the presence of random static disorder. The author has also applied the approach used here to nucleation [15], and this reference may be consulted for more details of the analysis performed below. The analysis required for nucleation is very similar to that required for our model of unfolding.

Different organisms have different genomes and so different sets of proteins, etc., inside their cells. Supattapone and coworkers [6] found that RNA molecules from mammals accelerated the protein conformational conversion whereas RNA from invertebrates did not. Thus, we

would like to model and try to understand species-to-species variability. To do so, we simply assume that the surfaces in different species are uncorrelated, then two species are modelled by two uncorrelated realisations of the surface. Of course, the surfaces present in closely related species in particular will be correlated due to their similar genomes, but we will leave the introduction of such correlations to future work.

Continuing, as only hydrophobic monomers interact, n_i is a sum of B independent random variables that are 1 with probability h and 0 with probability $1-h$. So, the probability distribution function of n_i, $p(n_i)$, is

$$p(n_i) = \frac{B!}{n_i!(B-n_i)!} h^{n_i}(1-h)^{B-n_i} \simeq \frac{\exp\left[-(n_i-m)^2/(2w^2)\right]}{(2\pi w^2)^{1/2}}\,, \tag{3}$$

where we have indicated that $p(n_i)$ is approximately a Gaussian for large B and n_i. $m = Bh$ is the mean, and the variance $w^2 = Bh(1-h)$. From now on we will neglect any deviations from the simple Gaussian distribution function of eq. (3) and the discrete nature of n_i and use this equation for $p(n_i)$.

Having chosen to model different species by uncorrelated realisations, we will examine fluctuations of the rate R between different realisations. We assume this variation between realisations is a reasonable model for variations between species. Returning to eq. (2) for the rate, using eq. (1) we obtain

$$R = N_{\rm s}^{-1}\nu \exp\left[-\Delta F_0^*\right] \sum_i^{N_{\rm s}} \exp\left[n_i\epsilon\right], \tag{4}$$

where the n_i are taken to be random variables drawn from the Gaussian distribution eq. (3). Except for constant factors, the rate R is equivalent to the partition function of the Random Energy Model (REM) of Derrida [16, 17]. The REM is a simple and well-studied model of glasses and other disordered systems.

Just as the average partition function of the REM can be obtained, we can obtain the average of the rate R,

$$\langle R\rangle = N_{\rm s}^{-1}\nu \exp\left[-\Delta F_0^*\right] \left\langle \sum_{i=1}^{N_{\rm s}} \exp\left[n_i\epsilon\right] \right\rangle \tag{5}$$

$$= \nu \exp\left[-\Delta F_0^*\right] \exp\left[m\epsilon + \epsilon^2 w^2/2\right]. \tag{6}$$

This is the average of R over many different realisations of the surface, *i.e.*, many different sets of the $N_{\rm s}$ random variables n_i that define a surface. As R is a sum over random variables, it itself is a random variable. For the large values of $N_{\rm s}$ considered here, the rate R is either self-averaging or non–self-averaging. It is self-averaging if for almost all realisations the rate of unfolding R is close to $\langle R\rangle$, *i.e.*, if R is almost the same for almost all realisations. Then the right-hand side of eq. (5) will be a good approximation to the rate R of any realisation. If it is non–self-averaging, then the rate R differs appreciably from one realisation to another, the values of R have a large spread and eq. (5) is unlikely to provide a good approximation to the value of R for a randomly selected realisation. R is non–self-averaging if and only if the sum of eq. (4) is dominated by one or a few terms: the variation comes from variation in the values of the largest terms in the sum. This is just as in the REM, see ref. [16] for details.

Recall that we are assuming that a realisation corresponds to a species. Thus, if R is self-averaging, then our model predicts that the rate of unfolding of a particular protein is almost the same in all or almost all species, whereas if it is not self-averaging then the rate of unfolding of a specific protein will vary significantly from one species to another.

We will now determine the boundary where the rate R crosses over from self-averaging to non–self-averaging. From eq. (4) we see that the rate R is dominated by unfolding configurations with values of n_i where the product of the number of configurations and $\exp[n_i\epsilon]$, is a maximum. The number of configurations is simply proportional to the probability of eq. (3). The maximum of the product $p(n_i)\exp[n_i\epsilon]$ is at $n_{\max} = m + \epsilon w^2$. Now, the *average* number of configurations around this value of n_i is just $N_{\rm s}p(n_{\max})$, and because this average is a sum over independent random variables (the n_i) the ratio of the fluctuations to the mean scales as $[N_{\rm s}p(n_{\max})]^{-1/2}$. Thus the fluctuations in the number of configurations that contribute the dominant amount to the rate, and hence the fluctuations in the rate itself are small relative to the mean if and only if $N_{\rm s}p(n_{\max}) \gg 1$. This is true whenever $2\ln N_{\rm s} - \epsilon^2 w^2 > 0$.

Thus, the boundary between self-averaging and non–self-averaging regimes is given by the equation

$$2\ln N_{\rm s} - \epsilon^2 w^2 = 0. \tag{7}$$

Note that $\epsilon^2 w^2$ is the variance of the distribution of interaction energies between the transition state and the surface. Thus the rate is self-averaging if and only if the logarithm of the number of possible configurations that the transition state can unfold in, is larger than half the variance of the interaction energy. This is the main result of this work. It is a very general result —it applies generally to activated processes in a random or near-random environment. Our conclusions here apply to any process with a rate given by an equation of the form of eq. (2), not just to protein unfolding *in vivo*. See ref. [15] for an application to nucleation at first-order phase transitions.

In the non–self-averaging regime, a single unfolding configuration can be responsible for a significant fraction of the entire rate of unfolding at the surface. This configuration must of course be the configuration with the largest value of n_i. We denote this largest value by x. If we define the probability distribution function, $p_{\rm ev}(x)$, of x, then the fraction of the rate R that is due to this extreme value is

$$f_{\rm ev} = \frac{\nu \exp\left[-\Delta F_0^*\right]}{N_{\rm s}\langle R\rangle} \int p_{\rm ev}(x) \exp\left[\epsilon x\right] {\rm d}x. \tag{8}$$

We can simplify eq. (8) by introducing the reduced variable $y = (x - m)/w$. Then, from eq. (8) and using eq. (5) for $\langle R\rangle$, we obtain

$$f_{\rm ev} = N_{\rm s}^{-1} \exp\left[-(\epsilon w)^2/2\right] \int {\rm d}y \exp\left[\epsilon w y\right] p_{\rm ev}(y), \tag{9}$$

where $p_{\rm ev}(y)$ is the probability distribution function for the maximum value of a set of $N_{\rm s}$ values taken from a Gaussian of zero mean and unit standard deviation. Note that although the absolute value of the rate R and of the contribution of the extreme value both depend on the mean m, $f_{\rm ev}$ does not. It depends only on the product ϵw, and on $N_{\rm s}$.

The determination of $p_{\rm ev}(y)$ is a standard problem in extreme-value statistics [18]. We start from the fact that the probability that the largest of $N_{\rm s}$ values is y is the probability that 1 of the $N_{\rm s}$ configurations has a value y, and all the remaining $N_{\rm s} - 1$ configurations have smaller values, multiplied by $N_{\rm s}$, as any one of the $N_{\rm s}$ configurations can have the largest value. Thus,

$$p_{\rm ev}(y) = N_{\rm s}p(y)p_<^{N_{\rm s}-1}(y), \tag{10}$$

where $p(y)$ is a normalised Gaussian of zero mean and unit standard deviation, and $p_<(y)$ ($p_>(y)$) is the probability of obtaining a number less (greater) than y from a Gaussian of zero mean and unit standard deviation. We are interested in the region where x is several standard

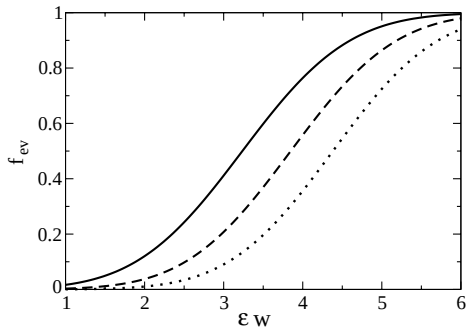

Fig. 2 – The mean fraction, $f_{\rm ev}$, of the rate R that is due to the configuration with the largest n_i, as a function of the product of the width of the Gaussian, w, and the interaction energy ϵ. The solid, dashed and dotted curves are for $N_{\rm s} = 1000$, 10000 and 100000 configurations, respectively.

deviations above the mean, $y \gg 1$. Now, $p_< = 1 - p_>$, and so for $y \gg 1$, $p_> \ll 1$, and we can rewrite eq. (10) as

$$p_{\rm ev}(y) \simeq N_{\rm s} p(y) \exp\left[-N_{\rm s} p_>(y)\right]. \tag{11}$$

Also, $p_>(y) = (1/2)\,{\rm erfc}(y/2^{1/2})$, which for $y \gg 1$ simplifies to $p_>(y) \simeq \exp[-y^2]/[(2\pi)^{1/2} y]$.

In fig. 2 we have plotted the fraction of the rate due to the configuration with the largest interaction energy, and so the lowest barrier, $f_{\rm ev}$, as a function of ϵw. We took $N_{\rm s} = 1000$, 10000 and 100000. Assuming that there are a few thousand different species inside a cell and that each can potentially interact with the transition state in a few ways, we end up with the estimate $N_{\rm s} \approx 10^4$. The other parameter is ϵw. The interaction strength of a pair of monomers is expected to lie in the range 1 to 3 (recall that ϵ is in units of $k_{\rm B}T$). If the fraction of hydrophobic monomers $h \approx 1/2$, then for $B \approx 5$ to 15 hydrophobic monomers, we have that $w \approx 1$ to 2. Combining these values for ϵ and w, we have that $\epsilon w \approx 0.5$ to 6. Returning to fig. 2, we see that as ϵw increases, so does $f_{\rm ev}$. For $N_{\rm s} = 10000$, eq. (7) is satisfied for $\epsilon w = 4.29$. For ϵw around this value, the configuration with the largest interaction energy already contributes a large amount to the total rate, on average. This large contribution will vary significantly from one realisation to the next, from one species to the next. So, the rate of unfolding of the protein will vary significantly from one species to the next, depending on whether the species has some part of a protein, RNA molecule, etc., that binds to the transition state unusually strongly. Our estimate for the possible values of ϵw *in vivo* goes up to around 6, so we estimate that the variation in the interaction free energies with a transition state may be large enough to cause random species-to-species variation. The RNA molecule or molecules found to catalyse the conversion is within our model the origin of one of the configurations that are outliers of the distribution, that interact most strongly with the transition state. Of course, if ϵw is small then the rate R has significant contributions from many unfolding configurations and so varies weakly from species to species, essentially due to variations in the rate being averaged out in accordance with the central-limit theorem.

In conclusion, Supattapone and coworkers [6] have found that cell extracts of some species but not others accelerate the conversion of the prion protein to a protease-resistant form. This conformational change must involve partial unfolding. Protein unfolding *in vivo* or in a cell extract occurs in a very complex and heterogeneous environment. There are a huge number of species present that potentially could interact with and stabilise the transition state of unfolding. A single strongly stabilising interaction could dramatically increase the rate of

unfolding. Here we have suggested a possible model for the species-to-species variation in the ability of cell extracts to accelerate prion protein conversion [6]. The model is a statistical one: interactions are modelled by random variables and different species by different uncorrelated realisations of the random interactions. We suggest that the acceleration is due to a strong interaction of the transition state for prion protein conversion with one or a few species of RNA molecules, and that this interaction is strong simply by chance. It is simply accidental that they reduce the free-energy barrier to unfolding. Proving this suggestion would require identifying the RNA molecule or molecules that interact with the prion protein and then demonstrating that there is no functional relationship between the protein and the RNA. Falsifying the suggestion is perhaps more straightforward, it only requires finding a functional relationship. The species-to-species variation then simply comes from the variation in the nucleotide sequences of RNA molecules from species to species. The RNA molecules that perform the same function in, say, mice and fruit flies, will have similar but not identical nucleotide sequences and so will have different interaction free energies with the transition state. Finally, it should be noted that it is also possible that the RNA molecule or molecules have evolved to interact with the prion protein, although we know of no evidence that they are under selection pressure to interact specifically with the transition state.

* * *

It is a pleasure to acknowledge that this work has benefited greatly from discussions with J. Cuesta. This work was supported by The Wellcome Trust (069242).

REFERENCES

[1] Dobson C. M., *Philos. Trans. R. Soc. B*, **356** (2001) 133.
[2] Harris D. A., *Clin. Micro. Rev.*, **12** (1999) 429.
[3] Aguzzi A. and Polymenidou M., *Cell*, **116** (2004) 313.
[4] Finkelstein A. V. and Ptitsyn O. G., *Protein Physics* (Academic Press, London) 2002.
[5] Creighton T. E., *Proteins: Structures and Molecular Properties* (Freeman, New York) 1993.
[6] Deleault N. R., Lucassen R. W. and Supattapone S., *Nature*, **425** (2003) 717.
[7] Wigner E., *SIAM Rev.*, **9** (1967) 1.
[8] Sear R. P. and Cuesta J. A., *Phys. Rev. Lett.*, **91** (2003) 245701.
[9] Cordeiro Y., Machado F., Juliano L., Aparecida Juliano M., Brentani R. R., Foguel D. and Silva J. L., *J. Biol. Chem.*, **276** (2001) 49400.
[10] Ryou C., Prusiner S. B. and Legname G., *J. Mol. Biol.*, **329** (2003) 323.
[11] Even *in vitro* protein unfolding is in general more complex than a simple crossing of a single barrier [4, 12, 13]. Also, it is far from clear that protein unfolding is the rate-limiting step in the progression of the formation of the protein fibrils in prion diseases. Here we simply assume that it is a simple activated process and explore the consequences.
[12] Onuchic J. N., Luthey-Schulten Z. and Wolynes P. G., *Annu. Rev. Phys. Chem.*, **48** (1997) 545.
[13] Pande V. S., Grosberg A. Yu. and Tanaka T., *Rev. Mod. Phys.*, **72** (2000) 259.
[14] Karpov V. G. and Oxtoby D. W., *Phys. Rev. B*, **54** (1996) 9734.
[15] Sear R. P., *Phys. Rev. E*, **70** (2004) 021605, cond-mat/0406019.
[16] Derrida B., *Phys. Rev. Lett.*, **45** (1980) 79; *Phys. Rev. B*, **24** (1981) 2613.
[17] The REM is used extensively in the study of protein folding, see, for example, refs. [12, 13]. But our use of the REM is rather different, the random variables here are interaction energies between the transition state and the surface, whereas in studies of protein folding they are the states of an isolated protein.
[18] Sornette D., *Critical Phenomena in Natural Sciences* (Springer-Verlag, Berlin) 2000.