http://dx.doi.org/10.5012/bkcs.2014.35.3.701

A Strategy for Searching a Rare Event 2[1]: Probability Distribution of Bindings[†]

Seung-Joon Jeon

Department of Chemistry, Korea University, Seoul 136-713, Korea. E-mail: sjjeon@korea.ac.kr
Received September 26, 2013, Accepted October 11, 2013

Key Words : Molecular searching, Facilitated diffusion, Bayesian inference, Posterior probability

In molecular biology, the central dogma has been the most important framework to explain major processes to maintain life. The first step is transcription in which the genetic information of a very tiny section of DNA is transferred to a piece of mRNA. During the process, several proteins are involved, like RNA polymerase and transcription factors (TFs). Especially, the binding of a TF to the right segment of DNA is critical, and has been studied a lot since early seventies.[2] For instance, a TF, called a *lac* repressor, makes gene expression regulate by site-specific binding to DNA, called a *lac* operon, in a living *E. coli* cell. The process is to search a rare event because a FT should find a right specific binding site among millions of sites on DNA.

Earlier, it was reported that searching time in an experiment *in vitro* was around two orders of magnitude faster than the Smoluchowsky diffusion limit process.[3] Since then, similar experimental results were interpreted by the facilitated diffusion(FD), which is a searching mechanism of 3D non-specific binding with 1D sliding along DNA.[2] Recently the FD mechanism has been supported by a few of single-molecule experiments, but also has been questioned due to the long sliding distance through crowding in the cytoplasm.[4,5] In addition there are a variety of controversial issues going on about the detailed mechanisms like non-specific binding and unbinding, hopping or sliding for 1D searching, and so on. They were usually focused on binding steps between a searching molecule and target sites. Previously I reported a qualitative expectation that unbinding steps might be more important than binding steps in searching a rear event through a probability theory called Bayes' theorem.[1] This report shows that probability distributions of bindings may also play a key role for a rare event searching.

The previous report[1] showed that the probability of searching a rare event, like binding TFs to the repressor on chromosomal DNA, depends more on detaching process rather than binding process of a right targeting. As the previous report, 'B' denotes the events of a searcher binding with a right targeting site, 'not B' with the other wrong sites. 'A' denotes the events that a searcher bound with any targeting site keeps binding with the site, and 'not A' means falling apart immediately after binding. P(A), P(not A), P(B), and P(not B) are their probabilities respectively. According to Bayes' theorem,[6] the conditional probability of a right targeting site given the events kept binding is as follows;

$$P(B|A) = \frac{P(B)\ P(A|B)}{P(A)}$$

$$(P(A) = P(B)P(A|B) + P(\text{not B})P(A|\text{not B}),\quad P(B) + P(\text{not B}) = 1) \tag{1}$$

, where P(A|B) is the conditional probability of events kept binding given a right targeting, P(A| not B) is the conditional probability of events kept binding given a wrong targeting.

Generally, a searcher is looking for a targeting site randomly. Therefore a searcher visits a lot of sites before a right targeting. Assume that a searcher keeps binding with a site on the *k*th visit. It should repeat attaching, detaching and re-attaching *k*-1 times before binding on the site of *k*th visit among total N decoy sites visitings. The conditional probability of a right targeting site given the events kept binding on the *k*th visit is similar to the Eq. (1) as follows;

$$P(B|A_k) = \frac{P(B)\ P(A_k|B)}{P(B)\ P(A_k|B) + P(\text{not B})P(A_k|\text{not B})} \tag{2}$$

According to Bayesian inference,[6] the posterior probability $P(B|A_k)$ can be inferred from a prior probability P(B), a likelihood function $P(A_k|B)$, and a marginal likelihood function which is the denominator of the Eq. (2), $P(A_k) = P(B)P(A_k|B) + P(\text{not B})P(A_k|\text{not B})$. As a crude assumption, the conditional probabilities of events kept binding given a targeting on the *i*th visit, $P(A_i|B)$ and $P(A_i|\text{not B})$ are same in any *i*th site visiting among total N sites visiting. Let them $P(A_i|B) = x$, $P(A_i|\text{not B}) = y$ in any *i*. Therefore $P(\text{not } A_i|B) = 1-x$, $P(\text{not } A_i|\text{not B}) = 1-y$. Assume that the searcher makes binding the right targeting site but detaching m times before the site of *k*th visit which keeps binding on the site. Then,

$$P(A_k|B) = (1-y)^{k-m-1} \cdot (1-x)^m \cdot x \tag{3}$$

$$P(A_k|\text{not B}) = (1-y)^{k-m-1} \cdot (1-x)^m \cdot y \tag{4}$$

$$P(A_k) = \frac{1}{N}(1-y)^{k-m-1} \cdot (1-x)^m \cdot x + \frac{N-1}{N}(1-y)^{k-m-1} \cdot (1-x)^m \cdot y \tag{5}$$

With the Eqs. (3), (4), (5), and P(B) = 1/N, P(not B) = (N–1)/N, the Eq. (2) will be simplified as follows;

$$P(B|A_k) = \frac{x}{x+(N-1)\cdot y} \tag{6}$$

[†]This paper is to commemorate Professor Myung Soo Kim's honourable retirement.

The always right targeting on *k*th visit means that the conditional probability of a right targeting site given the events kept binding on the *k*th visit, $P(B|A_k)$, should be 1, which means that there must be N=1 or y=0. N is a large number around 10^6-10^9, so y, which is the conditional probability of events kept binding given a wrong targeting on *k*th visit $P(A_k|\text{not B})$, should be zero. In other words, $P(\text{not } A_i|\text{not B})$ should be close to 1, no matter what value of $P(A_i|B)$. This result means that for searching a rare event detaching for any wrong sites is much more important than keeping binding for a right targeting site, as the previous report.

The above result is based on the crude assumption that the conditional probabilities of events kept binding given a targeting on *i*th visit can have only two values depending on right targeting or wrong targeting. In searching process, there should be diverse binding energies between a searcher molecule and sites on DNA, like an energy landscape model. Also keeping attaching on binding sites or detaching depends on thermal energy of the surroundings. Therefore, more realistically we can imagine that the conditional probabilities of $P(A_i|B)$ or $P(A_i|\text{not B})$ show a kind of continuous distribution. Assume that their distributions are normal distributions with their means μ_B, μ_{nB} and standard deviation σ_B, σ_{nB} respectively. The equation of normal distribution function is given by

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{7}$$

Bayes' theorem can also be applied in continuous distribution functions of the conditional probabilities,[6] so the Eq. (1) will be

$$P(B|A) = \frac{P(B)p(B|A)}{P(B)p(B|A) + P(\text{not B})p(A|\text{not B})} \tag{8}$$

where $p(A|B)$ and $p(A|\text{not B})$ are the density functions of the normal distribution for the conditional probabilities. Also the conditional probability of a wrong targeting site given the events kept binding is

$$P(\text{not B}|A) = \frac{P(\text{not B})p(A|\text{not B})}{P(B)p(A|B) + P(\text{not B})p(A|\text{not B})} \tag{9}$$

In order for a searcher to keep binding always on a right target, there must be $P(B|A) > P(\text{not B}|A)$. Using the Eqs. (8) and (9), we obtain

$$P(B)p(A|B) > P(\text{not B})p(A|\text{notB}) \tag{10}$$

This inequality is always valid if $p(A|\text{notB})=0$, which is the above result. Otherwise the inequality will be satisfied in a limited region of x. Using the Eq. (7), the inequality is

$$P(B)\frac{1}{\sigma_B\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_B}{\sigma_B}\right)^2} > P(\text{not B})\frac{1}{\sigma_{nB}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_{nB}}{\sigma_{nB}}\right)^2} \tag{11}$$

where x are probabilities between 0 and 1. So the general solution of the region of x, where $\sigma_B < \sigma_{nB}$, due to a narrow distribution of a right targeting, is

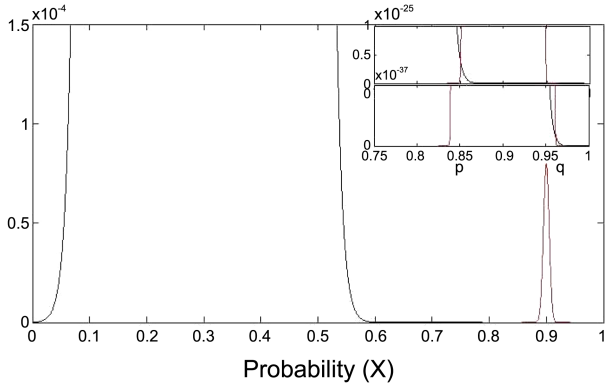

Figure 1. The distribution functions of P(not B)p(A|not B)(black) and P(B)p(A|B)(Red). p and q are the x values of the crossing points of the two functions, which are shown in the two insets.

$$\frac{-b-\sqrt{b^2-4ac}}{2a} < x < \frac{-b+\sqrt{b^2-4ac}}{2a} \tag{12}$$

where, $a = \sigma_B^2 - \sigma_{nB}^2$, $b = 2(\sigma_{nB}^2\mu_B - \sigma_B^2\mu_{nB})$

$$c = (\sigma_{nB}^2\mu_B^2 - \sigma_B^2\mu_{nB}^2) - 2\sigma_B^2\sigma_{nB}^2 ln\frac{p(B)\sigma_{nB}}{P(\text{not B})\sigma_B}$$

In binding of a TF on DNA, approximately $P(B) \approx 10^{-6}$, $P(\text{not B}) \approx 1$, and $\mu_B > \mu_{nB}$. Let's assume $\mu_B = 0.9$, $\mu_{nB} = 0.3$, $\sigma_B = 0.03$, $\sigma_{nB} = 0.05$. Figure 1 shows the posterior probability distribution function curves of P(B|A) and P(not B|A). The inequality is valid at the region $p < x < q$. This results indicates that the prior probability P(B) for a right targeting site, which is very small around 10^{-6}, becomes the posterior probability P(B|A) at the right targeting, which is always larger than P(not B|A) at a region of x. A searcher, like a TF, should keep binding only at a right site on the certain condition, if the likelihood probabilities show continuous distributions. The meaning of the range of x is still unclear, so required for further study. However it is speculated that one of the reasons for searching a rare event to maintain life might be probability distributions from diverse binding energies and thermal energy of the surroundings around binding sites in cell.

Acknowledgments. This paper dedicates to Professor Myung Soo Kim on the occasion of his honourable retirement. The author acknowledges to the support of Korea University on sabbatical leave for this work.

References

1. Jeon, S.-J. *Bull. Korean Chem. Soc.* **2012**, 33, 31. The first paper of the series of these works.
2. Xie, X. S.; Choi, P. J.; Li, G. W.; Lee, N. K.; Lia, G. *Annu. Rev. Biophys.* **2008**, *37*, 414.
3. Riggs, A. D.; Bourgeois, S.; Cohn, M. *J. Mol. Biol.* **1970**, *53*, 401.
4. Hammar, P.; Leroy, P.; Mahmutovic, A.; Marklund, E. G.; Berg, O. G.; Elf, J. *Science* **2012**, *336*, 1595.
5. Friedman, L. J.; Mumm, J. P.; Gelles, J. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 9740.
6. Link, W. A.; Barker, R. J. *Bayesian Inference*; Academic Press: San Diego, USA, 2010.