17th International Symposium on Mathematical and Computational Biology
Oct. 30 - Nov. 03 2017
Moscow, Institute of Numerical Mathematics, Russia

TOWARDS A THERMODYNAMICAL APPROACH OF PROTEIN DATABASES

R. P. Mondaini

S. C. de Albuquerque Neto

Federal University of Rio de Janeiro, RJ, Brazil
Centre of Technology, COPPE

OCTOBER 30, 2017

"Le savant doit ordonner ; on fait la science avec des faits comme une maison avec des pierres ; mais une accumulation de faits n'est pas plus une science qu'un tas de pierres n'est une maison.

The Scientist must set in order. Science is built up with facts, as a house is with stones. But a collection of facts is no more a science than a heap of stones is a house."

— Henri Poincaré, Science and Hypothesis

A derivation of the probability density function (pdf) for describing the formation of protein families through the solution of Fokker-Planck equations and the maximization of selected entropy measures. The present approach introduces a new statistical treatment of the formation and evolution of protein families and helps to support the classification of proteins into families and clans by methods of Statistical Mechanics.

Astronomical Almanacs (Ephemerides) / Biological Almanacs

Organization of the Sample Space

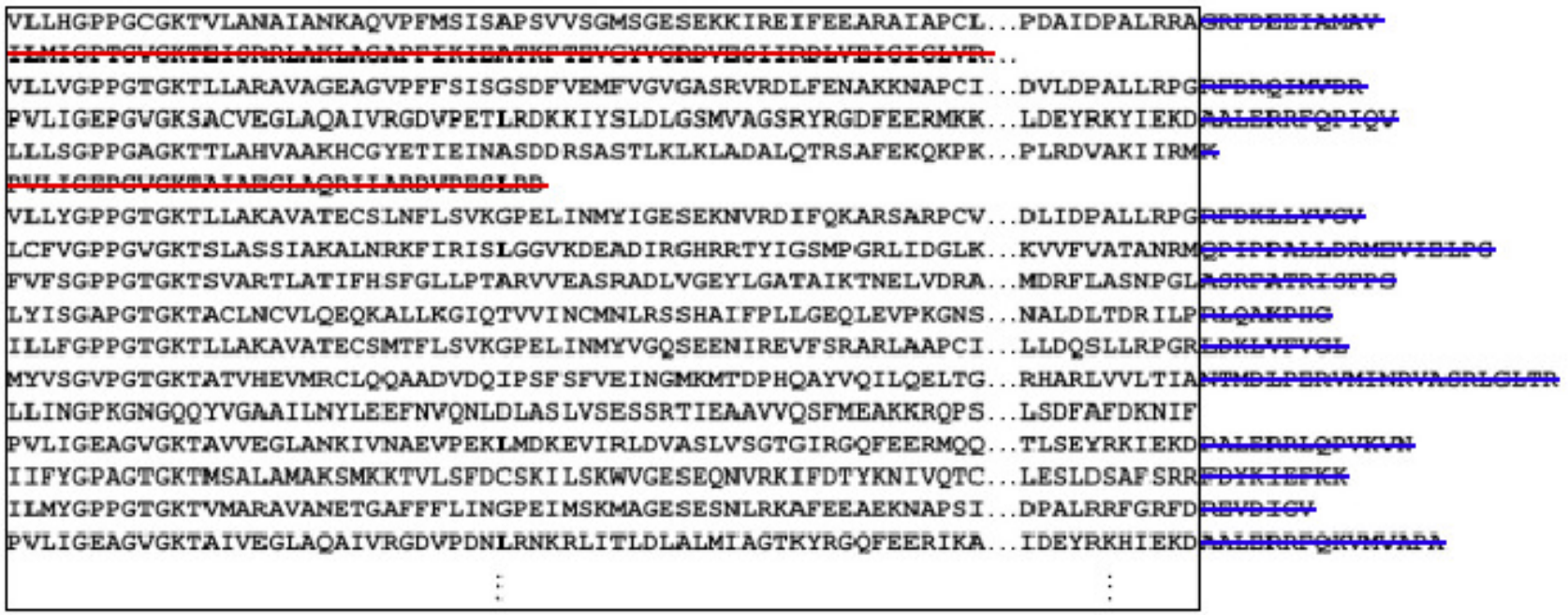

Figure: $(m \times n)$ block of amino acids — a representative of a protein family — an element of the sample space.

Each row has n_l amino acids, $l = 1, 2, \ldots, m$. We delete all domains such that $n_l < n$ and we also delete $(n_l - n)$ amino acids on all other domains.

The Ribosome Factory and the Organization of a Protein Family:

An example with $\Big((m = 6)\mathrm{X}(n = 9)\Big)$ and 12 different amino acids: V, P, L, C, F, H, I, Y, G, E, T, A.

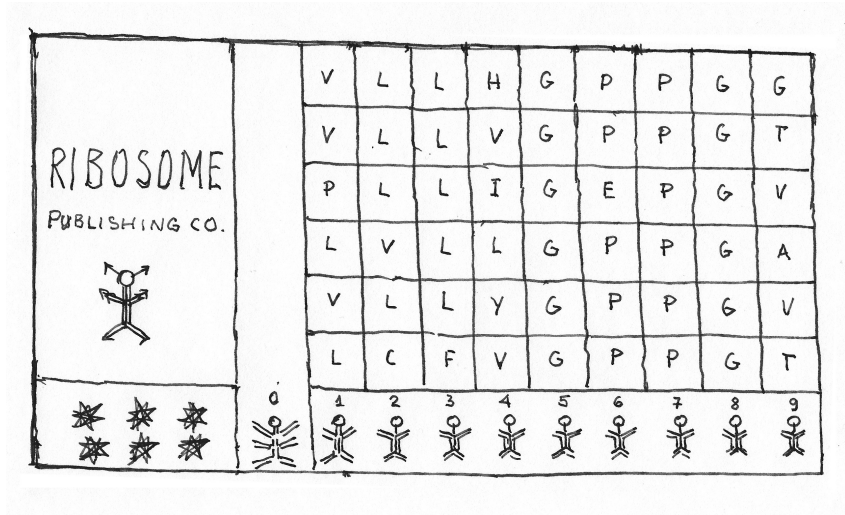

Figure: The Ribosome Publishing Co.

The "Ribosome Player" is tossing up simultaneously 6 dodecahedron dice in this example. After this first throw, his assistant will deliver the "amino acids books" to the 1st column. After the second throw, the first librarian will collect the books of his/her column and they are delivered to the librarian of the 2nd column. Meanwhile, the first librarian will allocate the amino acids of the second throw of his/her column and the second librarian will deliver his/her amino acids to the third librarian and so on.

Dice are assumed to be fair, but there is a previous game played with tetrahedra (faces A, C, G, U) and icosahedra to follow the guidelines of the Genetic Code and actually, a successful game for organizing protein families should take into consideration unfair icosahedron dice.

Random variables — 1) The probabilities of occurrence of amino acids in the columns of the $m \times n$ blocks.
n vectors of 20 components to be given by

$$p_j(a) = \frac{n_j(a)}{m}\,, \quad j = 1, 2, \ldots, n\,, \quad \sum_a p_j = 1$$

$a =$ A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y

$n_j(a)$ stands for the number of occurrences of the a-amino acid in the j^{th} column.

$$p_1 = \begin{pmatrix} p_1(A) \\ \vdots \\ p_1(Y) \end{pmatrix}\,, p_2 = \begin{pmatrix} p_2(A) \\ \vdots \\ p_2(Y) \end{pmatrix}\,, \ldots\,, p_n = \begin{pmatrix} p_n(A) \\ \vdots \\ p_n(Y) \end{pmatrix}$$

2) The joint probabilities of occurrence of amino acids — the ordered pair $\underline{ab}$ on an ordered pair of columns $\underline{jk} \longrightarrow \binom{n}{2} = \frac{n(n-1)}{2}$ square matrices p_{jk} of $(20)^2 = 400$ components

$$p_{jk} = \frac{n_j(a,b)}{m}\,, \qquad \begin{matrix} j = 1, 2, \ldots, n-1 \\ k = j+1, j+2, \ldots, n \end{matrix} \qquad , \quad \sum_{a,b} p_{jk}(a,b) = 1$$

$$a, b = \text{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}$$

$n_{jk}(a,b)$ stands for the number of occurrences of the ordered pair $\underline{ab}$ in columns $\underline{jk}$.

$$p_{jk} = \begin{pmatrix} p_{jk}(A,A) & \ldots & p_{jk}(A,Y) \\ \vdots & \ddots & \vdots \\ p_{jk}(Y,A) & \ldots & p_{jk}(Y,Y) \end{pmatrix}$$

These matrices can be arranged in a triangular array:

$$p = \begin{pmatrix} 0 & p_{12} & p_{13} & p_{14} & \cdots & p_{1\,n-2} & p_{1\,n-1} & p_{1\,n} \\ 0 & 0 & p_{23} & p_{24} & \cdots & p_{2\,n-2} & p_{2\,n-1} & p_{2\,n} \\ 0 & 0 & 0 & p_{34} & \cdots & p_{3\,n-2} & p_{3\,n-1} & p_{3\,n} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & p_{n-3\,n-2} & p_{n-3\,n-1} & p_{n-3\,n} \\ 0 & 0 & 0 & 0 & \cdots & 0 & p_{n-2\,n-1} & p_{n-2\,n} \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & p_{n-1\,n} \end{pmatrix}$$

3) Generalized joint probabilities of occurrences of amino acids — the ordered sequence of amino acids a_1, a_2, ..., a_{n-1}, a_n on an ordered set of columns j_1, j_2, ..., j_{n-1}, $j_n \longrightarrow \binom{n}{r}$ objects $p_{j_1 j_2 \ldots j_r}$ of $(20)^r$ components

$$p_{j_1 j_2 \ldots j_r}(a_1, a_2, \ldots, a_r) = \frac{n_{j_1 j_2 \ldots j_r}(a_1, a_2, \ldots, a_r)}{m}$$

$$\begin{aligned}
j_1 &= 1, 2, \ldots, (n - r + 1)\,; \\
j_2 &= (j + 1), (j + 2), \ldots, (n - r + 2)\,; \\
&\vdots \\
j_{r-1} &= (j_1 + r - 2), \ldots, (n - 1)\,; \\
j_r &= (j_1 + r - 1), \ldots, n
\end{aligned}$$

$$1 \leq r \leq n$$

$$a_1, a_2, \ldots, a_r = \text{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}$$

$$\sum_{a_1, a_2, \ldots, a_r} p_{j_1 j_2 \ldots j_r}(a_1, a_2, \ldots, a_r) = 1$$

Binomial Distribution $\longrightarrow$ Poisson Process

"Poissonization" $\longrightarrow$ current process in Statistics

Discrete binomial probability distribution with discrete time is "embedded" in a Poisson process of continuous time.

Let $p(n_j(t), t) \longrightarrow$ probability of finding n_j amino acids of the same kind in the j^{th} column. $\sigma \longrightarrow$ the probability per unit time that a transition of an amino acid from this column will occur. $\sigma\Delta t$ is the probability that the transition will occur in the interval of time Δt. We then have $(1 - \sigma\Delta t)$ as the probability that no transition will occur.

Master equation:

$$p(n_j(t + \Delta t), t + \Delta t) = \sigma\Delta t\, p(n_j(t) - 1, t) + (1 - \sigma\Delta t)\, p(n_j(t), t)$$

$$\Delta t \to 0 \longrightarrow \frac{\partial p(n_j(t), t)}{\partial t} = \sigma\left[p(n_j(t) - 1, t) - p(n_j(t), t)\right]$$

$$\frac{\partial p(n_0(t), t)}{\partial t} = -\sigma p(n_0(t), t)$$

At the "column $j = 0$" is the "Ribosome Factory". All amino acids are stored there before the start up of the process:

$$p(n_0(0), 0) = 1\,; \quad p(n_{j\neq 0}(0), 0) = 0$$

$$
\begin{aligned}
p(n_0(t),t) &= e^{-\sigma t} \\
p(n_1(t),t) &= e^{-\sigma t}\sigma t \\
p(n_2(t),t) &= e^{-\sigma t}\frac{(\sigma t)^2}{2} \\
p(n_2(t),t) &= e^{-\sigma t}\frac{(\sigma t)^3}{3!} \\
&\vdots \\
p(n_j(t),t) &= e^{-\sigma t}\frac{(\sigma t)^{n_j}}{n_j!}
\end{aligned}
$$

The moments are given by:

$$
\langle (n_j)^k \rangle = \sum_{n_j=0}^{\infty} n_j\, e^{-\sigma t}\frac{(\sigma t)^{n_j}}{n_j!} = e^{-\sigma t}\left(\sigma t\frac{\partial}{\partial(\sigma t)}\right)^k (\sigma t\, e^{\sigma t})
$$

then

$$
\begin{aligned}
\langle n_j \rangle &= \sigma t \\
\langle n_j^2 \rangle &= \sigma t + (\sigma t)^2
\end{aligned}
$$

Multinomial Distribution $\longrightarrow$ Maximization of Gibbs - Shannon entropy measure

Probability of observing all possible configurations of m amino acids in a column is found to be:

$$p(n_j(A), n_j(C), \ldots, n_j(Y)) = \frac{m!}{(20)^m \, n_j(A)! \, n_j(C)! \ldots n_j(Y)!}$$

$$S_j = -\sum_a \frac{n_j(a)}{m} \log \frac{n_j(a)}{m} = -\sum_a p_j(a) \log p_j(a)$$

For $m \gg 1$, $n \gg 1$:

$$\frac{1}{m} \log(p)_{\max} = 0 = -\log 20 + (S_j)_{\max}$$

Constrained Maximization

$$1 = \int_0^\infty p(n_j(t), t) \mathrm{d}n_j(t)$$

$$\langle n_j(t) \rangle = \int_0^\infty n_j(t) p(n_j(t), t) \mathrm{d}n_j(t)$$

$$\langle n_j^2(t) \rangle = \int_0^\infty n_j^2(t) p(n_j(t), t) \mathrm{d}n_j(t)$$

$$0 = \delta \int_0^\infty \Big[- p(n_j(t), t) \log p(n_j(t), t) - \lambda_0(t) p(n_j(t), t) - \lambda_1(t) n_j(t) p(n_j(t), t) - \lambda_2(t) n_j^2(t) p(n_j(t), t) \Big] \mathrm{d}n_j(t)$$

$$0 = \int_0^\infty \Big[- \log p(n_j(t), t) - 1 - \lambda_0(t) - \lambda_1(t) n_j(t) - \lambda_2(t) n_j^2(t) \Big] \delta p(n_j(t), t) \mathrm{d}n_j(t)$$

$$\therefore \; p(n_j(t), t) = \frac{1}{Z} e^{-\lambda_1(t) n_j(t) - \lambda_2(t) n_j^2(t)}$$

Partition Function:

$$Z = e^{1+\lambda_0(t)} = \int_0^\infty e^{-\lambda_1(t)n_j(t)-\lambda_2(t)n_j^2(t)}\mathrm{d}n_j(t) = \frac{1}{2\sqrt{\lambda_2(t)}\,M\big(y(t)\big)}$$

where:

$$M\big(y(t)\big) = \frac{1}{\sqrt{\pi}} \cdot \frac{e^{-y^2(t)}}{1-\mathsf{erf}\big(y(t)\big)}$$

$$\mathsf{erf}(y) = \frac{2}{\sqrt{\pi}}\int_0^y e^{-z^2}\mathrm{d}z$$

$$y = \frac{\lambda_1(t)}{2\sqrt{\lambda_2(t)}}$$

$$p(n_j(t),t) = 2\sqrt{\lambda_2(t)}\, M\Big(y(t)\Big)\, e^{-\lambda_1 n_j(t) - \lambda_2 n_j^2(t)}$$

$$\langle n_j(t)\rangle = -\frac{1}{\sqrt{\lambda_2(t)}}\Big(y(t) - M\Big(y(t)\Big)\Big)$$

$$\langle n_j^2(t)\rangle = \frac{1}{\lambda_2(t)}\left(\frac{1}{2} + y^2(t) - y(t)M\Big(y(t)\Big)\right)$$

$$M'(y) = -2yM(y) + M^2(y)$$

$$M''(y) = -2(1-2y^2)M(y) - 6yM^2(y) + 2M^3(y)$$

Asymptotic expansions: $0 < y \ll 1\,,\; M(y) \approx \dfrac{1}{\sqrt{\pi}}\left(1+\dfrac{2}{\sqrt{\pi}}\right) y$

$$y \gg 1\,,\; M(y) \approx y$$

$$\langle n_j(t)\rangle = -\frac{\partial \log Z}{\partial \lambda_1(t)} = \frac{1}{2\sqrt{\lambda_2(t)}}\frac{M'(y)}{M(y)}$$

$$\langle n_j^2(t)\rangle = \frac{1}{Z}\frac{\partial^2 Z}{\partial \lambda_1(t)} = -\frac{1}{4\lambda_2(t)}\left(\frac{M''(y)}{M(y)} - 2\frac{M'^2(y)}{M^2(y)}\right)$$

$$F = \langle n_j(t)\rangle + \langle n_j^2(t)\rangle - \frac{1}{\lambda_1(t)}S$$

$$Z = e^{-\lambda_1 F}$$

The Fokker-Planck equation:

Explicit dependence of the probability with the number of particles.
For an amino acid to be transfered from the j^{th} column, the probability of this process can be written:

$$p(n_j(t)-1,t) = p(n_j(t),t) - \frac{\partial p(n_j(t),t)}{\partial n_j(t)} + \frac{1}{2}\frac{\partial^2 p(n_j(t),t)}{\partial n_j^2(t)}$$

We then have from the Master Equation

$$p(n_j(t+\Delta t),t+\Delta t) = \sigma\Delta t\, p(n_j(t)-1,t) + (1-\sigma\Delta t)\, p(n_j(t),t)\,,$$

$$\frac{\partial p(n_j(t),t)}{\partial t} + \sigma\frac{\partial p(n_j(t),t)}{\partial n_j(t)} - \frac{\sigma}{2}\frac{\partial^2 p(n_j(t),t)}{\partial n_j^2(t)} = 0$$

Integrating from 0 to ∞ and using the boundary conditions,

$$\lim_{n\to\infty}\big(n_j(t)\big)^K p(n_j(t),t) = 0\,; \quad \lim_{n\to 0} p(n_j(t),t) = 0.$$

$$\sigma t = -\frac{1}{\sqrt{\lambda_2(y)}}\big(y - M(y)\big)$$
$$\sigma t + \sigma^2 t^2 = \frac{1}{\lambda_2(y)}\left(\frac{1}{2} + y^2 - yM(y)\right)$$

We then look for a point $y = c$ in the parameter space such that $M^2(c) \approx 0$ and we consider the Taylor expansion:

$$\begin{aligned} M(y) \approx & M(c) + M'(c)(y-c) + \mathcal{O}(y-c)^2 \\ & = \big(1 - 2c(y-c)\big)M(c) + \mathcal{O}(y-c)^2 \end{aligned}$$

$$\frac{1}{\lambda_2(y)}\left(\frac{1}{2} + 2c^3 M(c) + (1-2c^2)M(c)y\right) = \sigma t$$
$$\frac{1}{\sqrt{\lambda_2(y)}}\left(y - M\Big(c\big(1 - 2c(y-c)\big)\Big)\right) = \sigma t$$
$$y = \frac{\lambda_1(t)}{2\sqrt{\lambda_2(y)}}$$

$$\lambda_1(y) = \frac{1 + 2(1 - 2y^2 + 2y^3)M(y)}{1 + 2yM(y)},\ \lambda_2(y) = \frac{1 + 2(1 - 2y^2 + 2y^3)M(y)}{2\sigma t}$$

$$M(y) \equiv \frac{1}{\sqrt{\pi}} \frac{e^{-y^2}}{1 - \mathsf{erf}(y)}$$

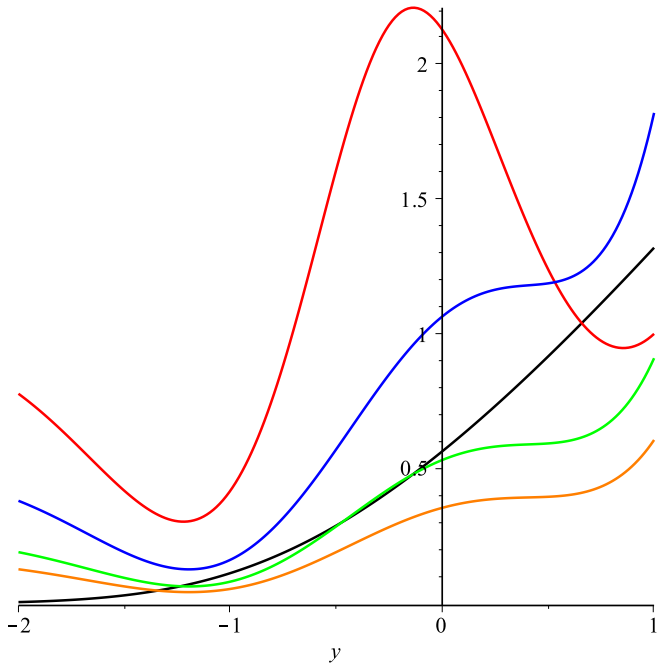

Figure: $\big(M(y)$ —$\big)$, $\big(\lambda_1(y)$ —$\big)$, $\big(\lambda_2(y), \sigma t = 1$ —$\big)$, $\big(\lambda_2(y), \sigma t = 2$ —$\big)$, $\big(\lambda_2(y), \sigma t = 3$ —$\big)$.

$$p(n_j(t),t) = 2\sqrt{\lambda_2(y)\,\pi}\;\frac{e^{-\lambda_2(y)\left(n_j(t)-\frac{\sigma t}{1+2yM(y)}\right)^2}}{1+\mathrm{erf}\left(\frac{\sigma t\sqrt{\lambda_2(y)}}{1+2yM(y)}\right)}$$

1) $p_1(n_j(t),t)\,,\quad M(y)\approx 0\,,\quad \lambda_2(y)=1\,,\quad y=-3\,,\quad \sigma t=3.$

2) $p_2(n_j(t),t)\,,\quad M(y)\approx 0.030245\,,\quad M^2(y)\approx 0\,,\ \lambda_2(y)=0.06333\,,\quad y=-1.5\,,\quad \sigma t=3.$

3) $p_3(n_j(t),t)\,,\quad M(y)\approx 0.053828\,,\quad M^2(y)\approx 0\,,\quad \lambda_2(y)=0.045123\,,\quad y=-1.3\,,\quad \sigma t=3.$

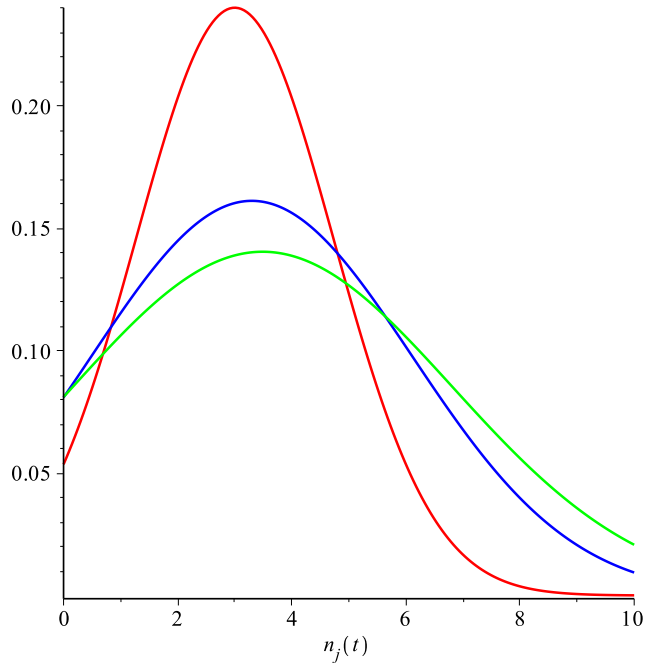

Figure: $\left(p_1(n_j(t),t)\ \text{—}\right)$, $\left(p_2(n_j(t),t)\ \text{—}\right)$, $\left(p_3(n_j(t),t)\ \text{—}\right)$.

References

1. R.P. Mondaini, S.C. de Albuquerque Neto — Entropy Measures and the Statistical Analysis of Protein Family Classification – BIOMAT 2015 (2016) 193–210.
2. R.P. Mondaini, S.C. de Albuquerque Neto — The Pattern Recognition of Probability Distributions of Amino Acids in Protein Families – BIOMAT 2016 (2017) 29–50.
3. R.P. Mondaini — A Survey of Geometric Techniques for Pattern Recognition of Probability of Occurrence of Amino Acids in Protein Families – BIOMAT 2016 (2017) 304–326.
4. R.P. Mondaini — Entropy Measures Based Method for the Classification of Protein Domains into Families and Clans – BIOMAT 2013 (2014) 209–218.
5. R.D. Finn et al. — Pfam: Clans, web tools and services – Nucleic Acids Research, 34 (2006) D247–D251.
6. M. Punta et al. — The Pfam Protein Families database – Nucleic Acids Research, 40 (2012) D290–D301.
7. R.D. Finn et al. — The Pfam Protein Families database – Nucleic Acids Research, 42 (2015) D222–D230.
8. R.D. Finn et al. — The Pfam Protein Families database – Nucleic Acids Research, 44 (2016) D279–D285.
9. Jan Hermans, Barry Lentz — Equilibria and Kinetics of Biological Macromocules, Wiley 2014, 441–447.
10. Yu.B. Rumer, M.Sh. Ryvkin — Thermodynamics, Statistical Physics and Kinetics – Mir Publishers 1980, pp. 576.
11. H. Risken — The Fokker-Planck Methods of Solutions and Applications 2nd edition, Springer Series in Synergetics, 2008.
12. N.G. Van Kampen — Stochastic Process in Physics and Chemistry, 3rd edition, North-Holland, 2007.
13. E.T. Jaynes — Probability Theory - The Logic of Science – Cambridge University Press, 2003.
14. M.H. DeGroot, M.J. Schervish — Probability and Statistics – Pearson Education 4th edition, 2012, pp. 597–604.
15. C. Beck — Generalized Information and Entropy Measures in Physics – Contemporary Physics 50 (4) (2009) 495–510.
16. L.R. Jaisingh — Statistics for the Utterly Confused – 2nd edition, McGraw-Hill, Publ. Inc., 2006
17. K. Sneppen, G. Zocchi – Physics in Molecular Biology, Cambridge Univ. Press, 2005.