

Sequence variation in ligand binding sites in proteins

Thomas J. Magliery^{1,2*} and Lynne Regan^{1,3*}

Departments of Molecular Biophysics & Biochemistry¹ and Chemistry³, New Haven, Connecticut 06520, U.S.A.

²Present Address: Departments of Chemistry and Biochemistry, The Ohio State University, 100 W. 18th Ave., Columbus, Ohio 43210, U.S.A.

*To whom correspondence should be addressed. TJM: E-mail magliery@chemistry.ohio-state.edu. Phone +1 (614) 247-8425. Fax +1 (614) 292-1685. LR: E-mail lynne.regan@yale.edu. Phone +1 (203) 432-9843. Fax +1 (203) 432-5767.

Supplementary Material

The relationship between multinomial probability and relative entropy can be demonstrated using the Stirling approximation for factorials:

$$\ln x! \approx \left(x + \frac{1}{2}\right) \ln x - x + \frac{1}{2} \ln 2\pi \approx x \ln x - x$$

Taking the natural log of the multinomial probability gives:

$$\ln P_{mult} = \ln N! - \ln n_1! - \ln n_2! \dots + n_1 \ln f_1 + n_2 \ln f_2 \dots$$

Substitution of the factorials with the Stirling approximation gives:

$$\ln P_{mult} = N \ln N - N - (n_1 \ln n_1 - n_1) \dots + n_1 \ln f_1 \dots$$

Since the sum of n_x over all x is N , these cancel to give:

$$\ln P_{mult} = N \ln N - \sum_x n_x \ln \frac{n_x}{f_x}$$

Subtracting $(n_x \ln N)$ from each term in the summation, which sums to $(N \ln N)$ over all x , gives:

$$\ln P_{mult} = - \sum_x n_x \ln \frac{p_x}{f_x}$$

Where $p_x = n_x/N$. Therefore,

$$\ln P_{mult} = -N \sum_x p_x \ln \frac{p_x}{f_x} = -N \times D(p \parallel f)$$

In other words, dividing the natural log of the multinomial probability by the total number of sequences gives the negative of the relative entropy.