Statistical distribution of isomerization energies in a protein ensemble

Giorgio F. Signorini

11th November 2003

Department of Chemistry, University of Florence
via della Lastruccia, 3 - 50019 Sesto Fiorentino - Italy
signo@chim.unifi.it

Consider an ensemble of proteins, each containing one or more instances of a certain group which can assume several forms (for example, histidine, where three protonation forms are known). For each protein, all possible isomers are combinations of forms of each group. The energy of an isomer can be expressed as the difference between its energy and the energy of a reference isomer for the same protein. The expected statistical distribution of this energy difference for the whole ensemble is shown in the case that the reference isomer is chosen at random for each protein.

1 Energy distribution in one protein

1.1 energy as the sum of independent variates

In a protein with N *independent* groups, the total energy is the sum of all group energies, plus the energy of the rest, E_0:

$$E = E_0 + E_1 + E_2 + \ldots + E_N \tag{1}$$

This may be taken as a definition of "independent group"; two groups whose energies are not additive should be considered as one big group[1]. If the statistical distribution of each group's energy (due to variations such as the displacement of one hydrogen) is $p_i(X)$, the distribution of total energy E is the *convolution of all these distributions*[2]. For example, if $N = 2$, and assuming $E_0 = 0$:

$$P(E) = \int_{-\infty}^{\infty} p_1(X) p_2(E - X) dX \tag{2}$$

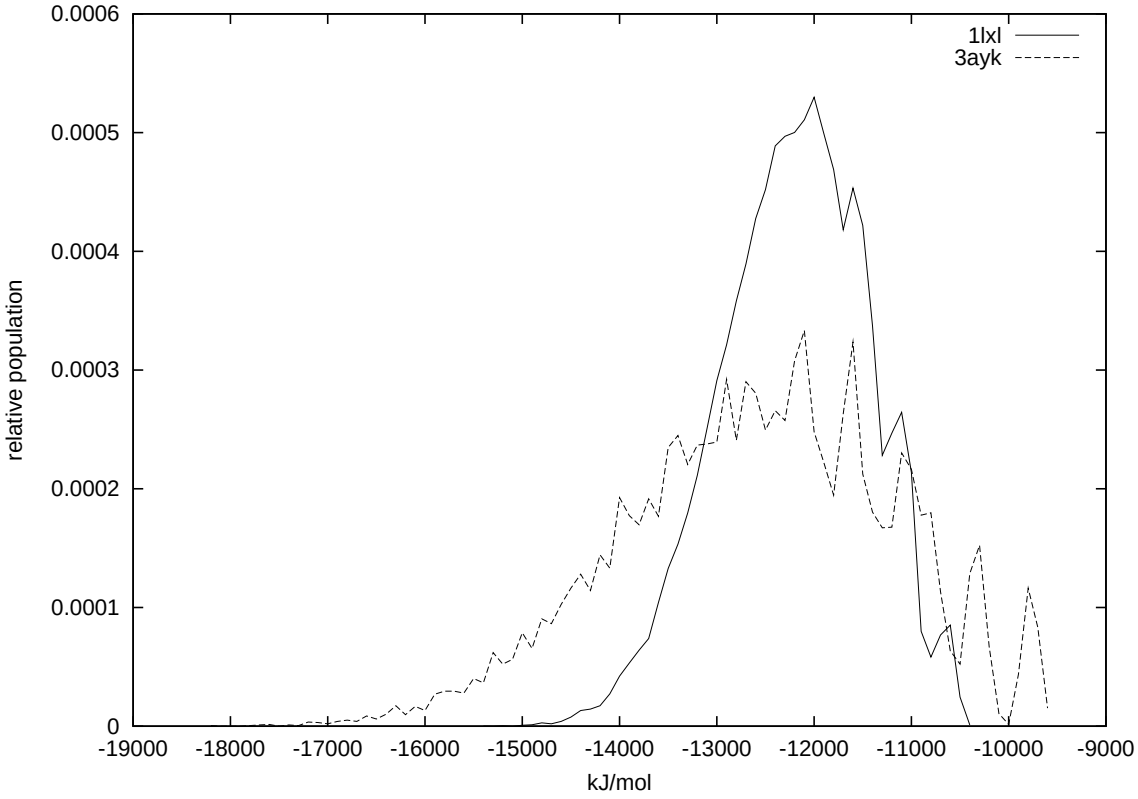
1.2 limit of the distribution

Following the Central Limit Theorem, (cf. for example [1]), the distribution of energies in one protein, $P(X)$, tends (rather quickly with increasing number of groups, N) to a Gaussian

[1]Note that in the accompanying paper, N indicates the total number of histidines, whether they are independent or not

[2]The equivalence between the distribution of a sum of variates and the convolution of single variate distributions is well known and illustrated e. g. in [1] (cap.16), both for continuous and discrete distributions.

whose centroid is the sum of the centroids of each group's distribution (and variance is the sum of variances). The following figure shows the distribution of energies due to isomerization of histidines in a protein with 9 histidines (`3ayk`) and in one with 10 histidines (`1lxl`) :



1.3 distribution of energy differences

The distribution of energy *differences* with respect to the energy of the isomer reported in the PDB, $X = E - E_{PDB}$, is simply

$$P'(X) = P(E) = P(X + E_{PDB}) \tag{3}$$

2 Global distribution of energy differences

2.1 global distribution

We now consider a collection of L proteins (in the case discussed in the accompanying paper, $L = 409$), where a reference isomer is defined for each protein; the distribution of energy differences for each protein is $P_n(X + X_n^0)$, where X_n^0 is the reference isomer of protein n.

We define the global distribution of X as the sum of the L single-protein distributions, giving each protein the same weight, $\frac{1}{L}$:

$$P(X;\, X_1^0 X_2^0 ...) = \frac{1}{L} \sum_n^L P_n(X + X_n^0) \tag{4}$$

with

$$\int_{-\infty}^{\infty} P_n(X) dX = 1 \tag{5}$$

The distribution for protein n may be written

$$P_n(X) = \frac{1}{M_n} \sum_i^{M_n} \delta(X - E_{ni}) \tag{6}$$

(where M_n is the number of isomers of protein n and E_{ni} is the energy of its i-th isomer), and the global distribution is

$$P(X) = \frac{1}{L} \sum_n^L \frac{1}{M_n} \sum_i^{M_n} \delta(X + X_n^0 - E_{ni}) \tag{7}$$

Noting that $P_n(X + X_n^0)$ always includes the trivial term $\frac{1}{M_n}\delta(X)$ corresponding to the case $E_{ni} = X_n^0$

$$P_n(X + X_n^0) = \frac{1}{M_n} \sum_i^{M_n}{}' \delta(X + X_n^0 - E_{ni}) + \frac{\delta(X)}{M_n} \tag{8}$$

(where the symbol $\sum'$ means that the term $E_{ni} = X_n^0$ is to be excluded from the sum), we may write the global distribution (7) as

$$P(X) = \frac{1}{L} \sum_n^L \frac{1}{M_n} \sum_i^{M_n}{}' \delta(X + X_n^0 - E_{ni}) + \delta(X) \frac{1}{L} \sum_n^L \frac{1}{M_n} \tag{9}$$

$$= p(X) + \delta(X) \frac{1}{L} \sum_n^L \frac{1}{M_n} \tag{10}$$

with

$$p(X) \equiv \frac{1}{L} \sum_n^L \frac{1}{M_n} \sum_i^{M_n}{}' \delta(X + X_n^0 - E_{ni}) \tag{11}$$

When representing real distributions through histograms, the pulse in the origin (second term in (10)) is unpractical (since the distribution is a density, the height of the pulse is inversely proportional to the channel width); it is often more useful to represent $p(X)$ than $P(X)$, keeping in mind, however, that the integral of $p(X)$ is $(1 - \frac{1}{L}\sum_n \frac{1}{M_n})$ and not 1.

2.2 average global distribution

When X_n^0 is chosen at random, the quantity of interest is the *average distribution* with respect to all possible choices of X_n^0, which is the sum of single-protein averages:

$$\bar{P}(X) = \frac{1}{L} \sum_n^L \bar{P}'_n(X) \tag{12}$$

The single-protein average $\bar{P}'_n(X)$ is obtained by multiplying the distribution with origin in X_n^0 by the probability, $P_n(X_n^0)$, that the origin will be placed in X_n^0, and integrating over all possible origins:

$$\bar{P}'_n(X) = \int_{-\infty}^{\infty} P_n(X_n^0) P_n(X + X_n^0) dX_n^0 \tag{13}$$

This expression for $\bar{P}'_n(X)$ shows that *the average distribution for a single protein is the autocorrelation function of* P_n:

$$\bar{P}'_n(X) = P_n \star P_n \tag{14}$$

It is well known that the distribution of the difference between two variates (in this case, $X = (X + X_n) - X_n$) is the correlation between the two distributions.

The average global distribution (12) is thus

$$\bar{P}(X) = \frac{1}{L} \sum_n^L P_n \star P_n \tag{15}$$

Recall that the autocorrelation is always an even function and has a maximum in $X = 0$. For a proof see for example ref. [1] (chap. 3, appendix).

As an example, consider the simple case of one protein with only two isomers, and thus two energies a e b:

$$P_n(X) = \frac{1}{2} \left[\delta(X - a) + \delta(X - b)\right] \tag{16}$$

The single-protein average is

$$\bar{P}'_n(X) = P_n \star P_n = \frac{1}{2}\delta(X) + \frac{1}{4} \left[\delta(X + (a - b)) + \delta(X - (a - b))\right] \tag{17}$$

which is simply a pulse in the origin plus two half-pulses at $(a-b)$ and $-(a-b)$. If all proteins in the ensemble are two-isomer proteins, the global distribution is the sum of many such functions, that is a pulse in 0 plus many small pulses scattered symmetrically around zero.

In the general case, P_n is given by (6), and

$$P_n \star P_n = \frac{1}{M_n^2} \sum_{i,j}^{M_n} \delta[X - (E_{ni} - E_{nj})] \tag{18}$$

$$\bar{P}(X) = \frac{1}{L} \sum_n^L \frac{1}{M_n^2} \sum_{i,j}^{M_n} \delta[X - (E_{ni} - E_{nj})] \tag{19}$$

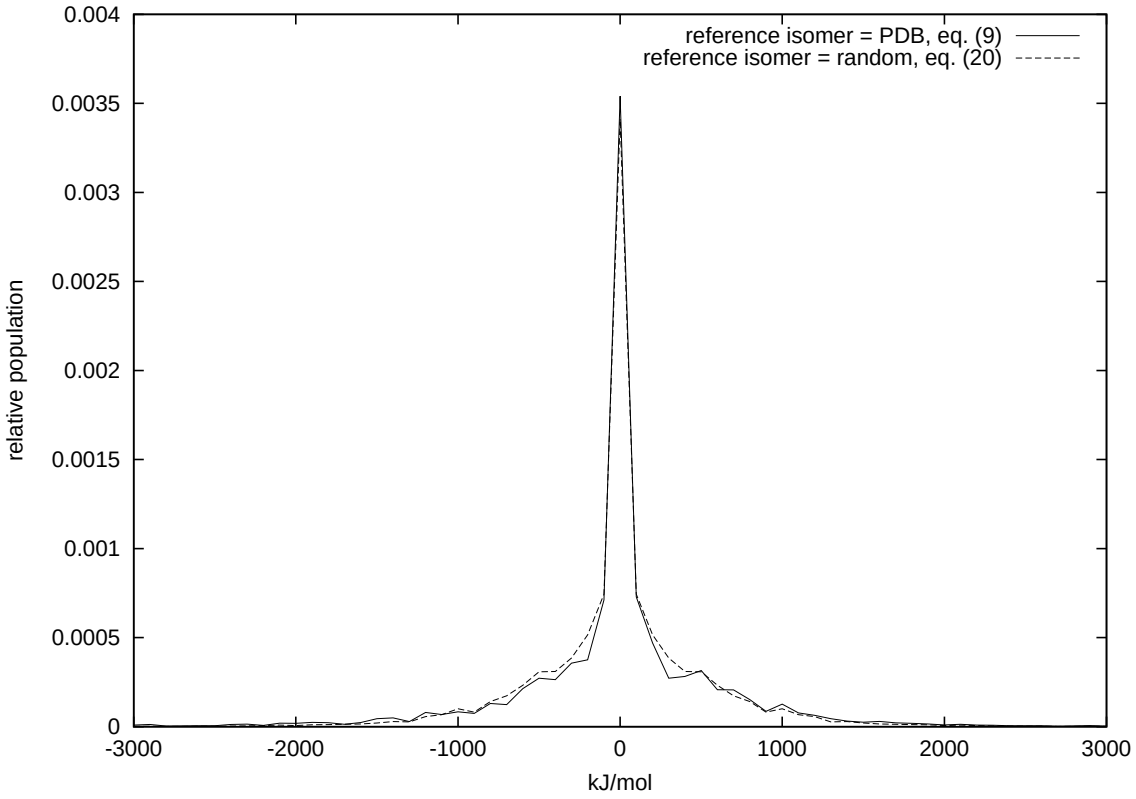
Here again we have a trivial term due to the cases $i = j$. One may single it out, similarly to (9):

$$\bar{P}(X) = \frac{1}{L} \sum_n^L \frac{1}{M_n^2} \sum_{i \neq j}^{M_n} \delta[X - (E_{ni} - E_{nj})] + \delta(X) \frac{1}{L} \sum_n^L \frac{1}{M_n} \tag{20}$$

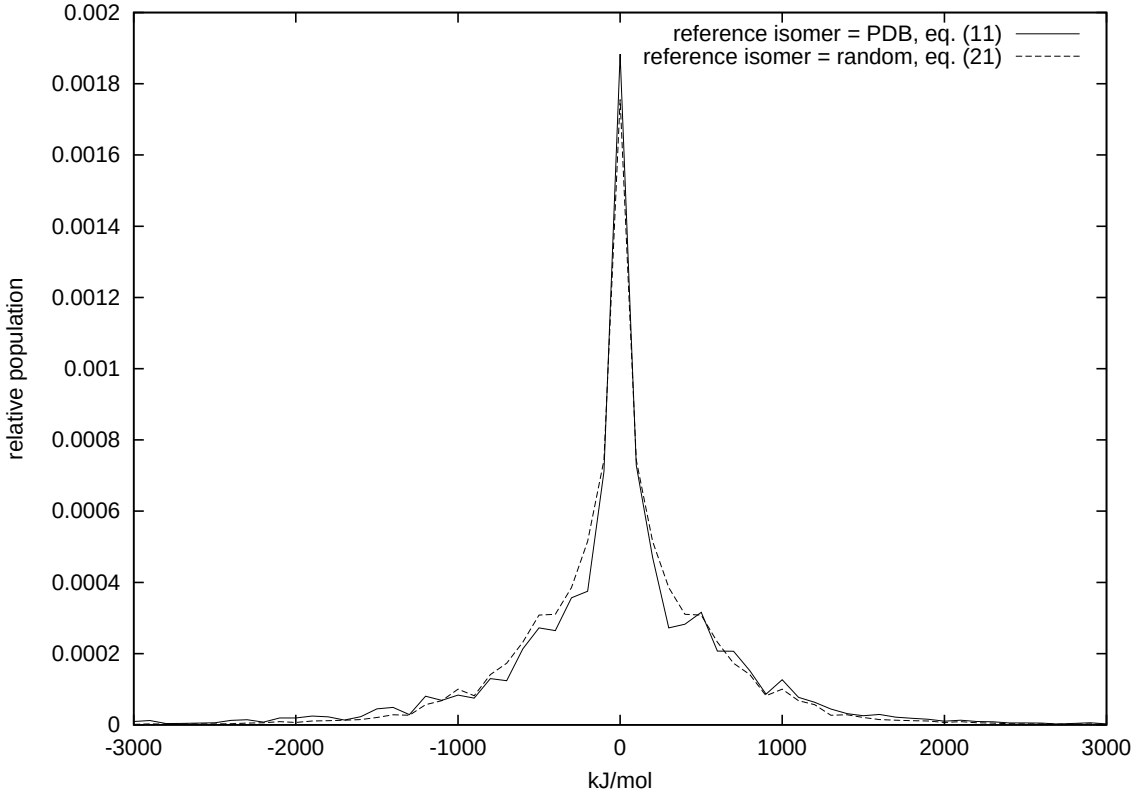
$$= \bar{p}(X) + \delta(X) \frac{1}{L} \sum_n^L \frac{1}{M_n} \tag{21}$$

(note that $\bar{p}(X)$ is exactly the average, with respect to all possible choices of X_n^0, of $p(X)$ as defined by (11), in the same way as $\bar{P}(X)$ in (19) is the average of (9))

In a collection of proteins whose energies E_{ni} are known, the average distribution $\bar{P}(X)$ can be calculated from (19). In the case reported in the accompanying paper, the following average distribution is obtained (note the comparison with the "observed" distribution, that is the distribution obtained when the reference isomer is the PDB one):



The similarity between the two distributions is even more clear if comparison is made between $p(X)$ and $\bar{p}(X)$, that is, if distributions are "cleaned" from the trivial term in 0:



2.3 models and approximations with analytical functions; distribution of σ

It may be useful to write an analytical expression for $\bar{P}(X)$ in model cases.

If the average distribution of each protein can be expressed by a function depending on only one parameter σ, (for example, a Gaussian of width σ centered in zero), $P(X;\sigma_n)$, the global distribution will be:

$$\bar{P}(X) = \frac{1}{L}\sum_{n}^{L} P(X;\sigma_n) \tag{22}$$

or, if the number of distributions in the interval from σ to $(\sigma + d\sigma)$ is $\rho(\sigma)d\sigma$, and σ ranges from σ_o to σ_1, it will be the integral:

$$\bar{P}(X) = \int_{\sigma_0}^{\sigma_1} \rho(\sigma)P(X;\sigma)d\sigma \tag{23}$$

with

$$\int_{\sigma_0}^{\sigma_1} \rho(\sigma)d\sigma = 1 \tag{24}$$

2.3.1 Models for $P(X;\sigma)$

If, in the first approximation, a *uniform* distribution for σ is assumed ($\rho(\sigma) = 1/(\sigma_1 - \sigma_0)$ if $\sigma \in [\sigma_0, \sigma_1]$, $\rho(\sigma) = 0$ otherwise), an analytical expression of $\bar{P}(X)$ can be devised in some model cases. For example:

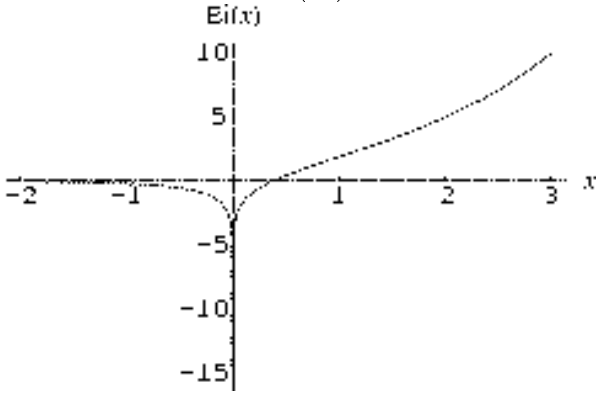
1. **if all proteins contain many groups**, each single-protein distribution tends to a Gaussian. The average distribution is this Gaussian's autocorrelation, that is again a Gaussian, centered in the origin and with the same variance. If the Gaussians' variances span the interval between σ_0 e σ_1 *uniformly*, then

$$\bar{P}(X) = \frac{1}{\sigma_1 - \sigma_0}\int_{\sigma_0}^{\sigma_1} \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{X^2}{2\sigma^2})d\sigma \tag{25}$$

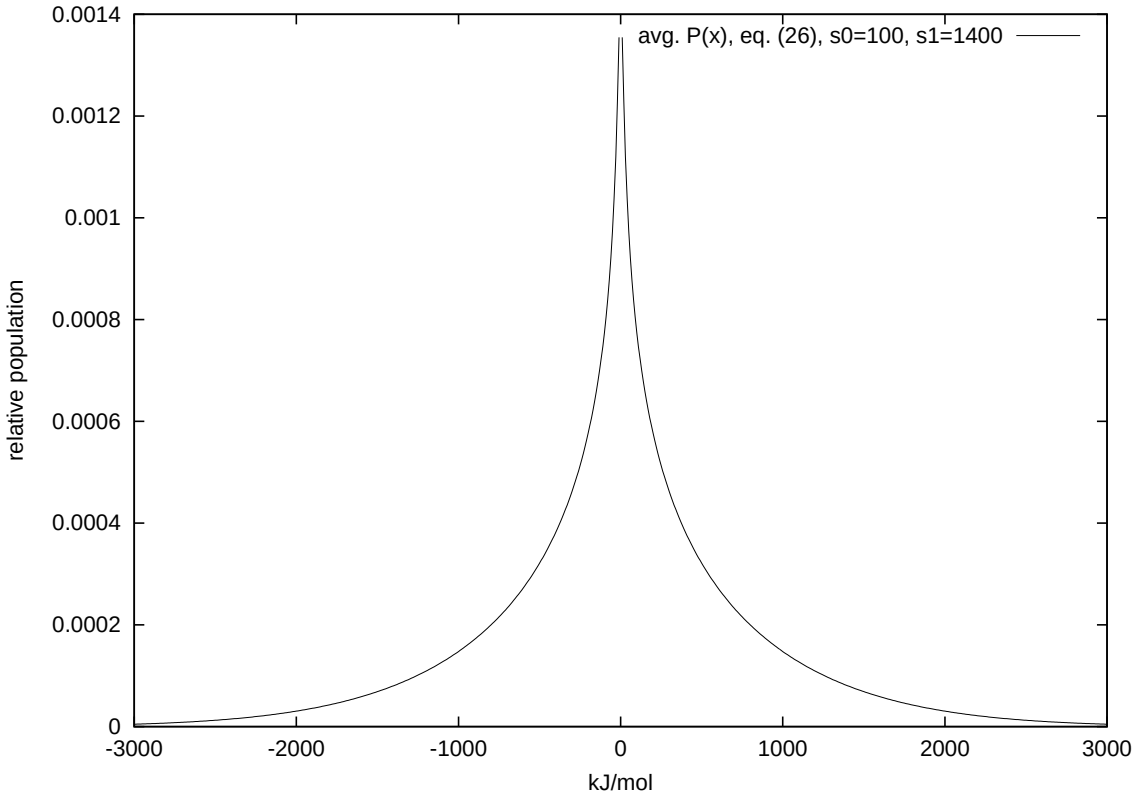
The value of this integral is

$$\bar{P}(X) = \frac{ei(-\frac{X^2}{2\sigma_0^2}) - ei(-\frac{X^2}{2\sigma_1^2})}{(\sigma_1 - \sigma_0)2\sqrt{2\pi}} \tag{26}$$

where the function $ei(X)$ is the so-called *exponential integral*:



Basically, the distribution has a "lambda" or "witch's hat" shape (cf. the *negative part* of the function plot, noting that the difference between the two functions in the upper half of the fraction in (26) is positive). The following is its value for $\sigma_0 = 10, \sigma_1 = 1400$ (these values refer to the observed distribution of σ, see below):

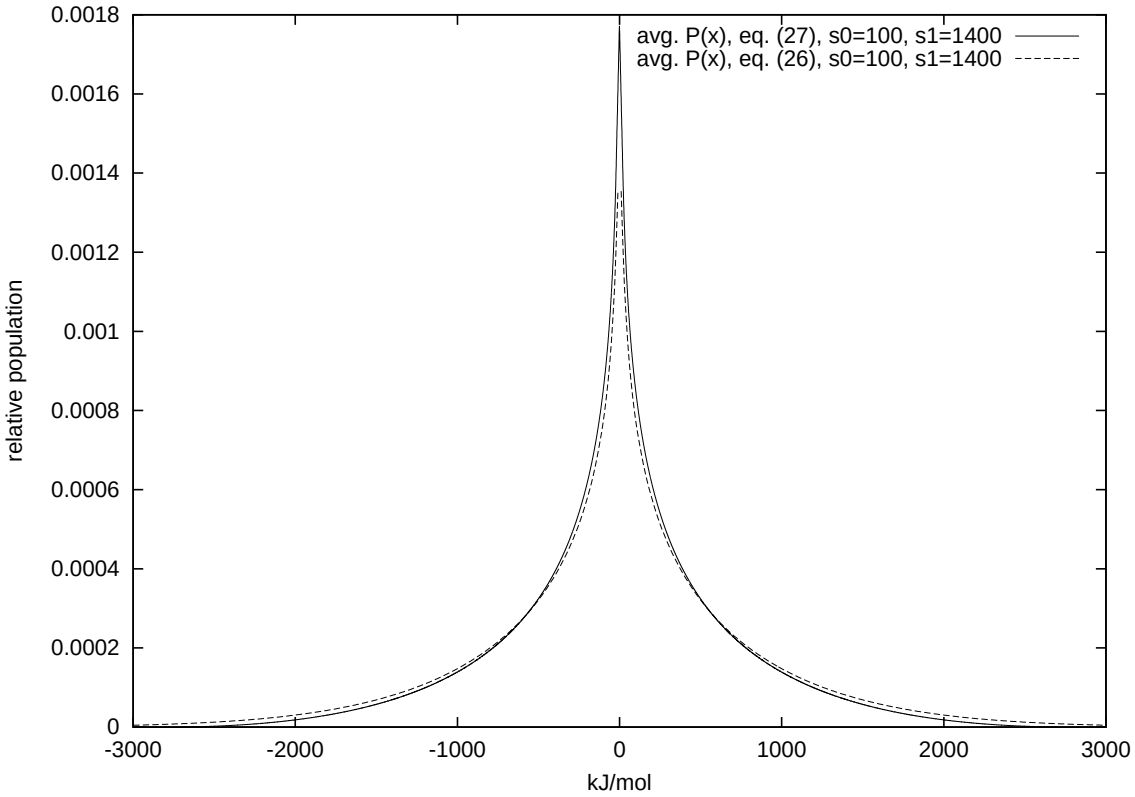


2. **if all proteins have few groups**, the single-protein distribution can be approximated by a "rectangular" distribution ($P(X;\sigma) = \frac{1}{2\sigma} \Leftrightarrow |X| \leq \sigma$), whose autocorrelation is "triangular" ($P(X;\sigma) = (2\sigma - |X|)/4\sigma^2 \Leftrightarrow |X| < 2\sigma$); keeping the assumption $\rho(\sigma) = 1/(\sigma_1 - \sigma_0)$ one obtains

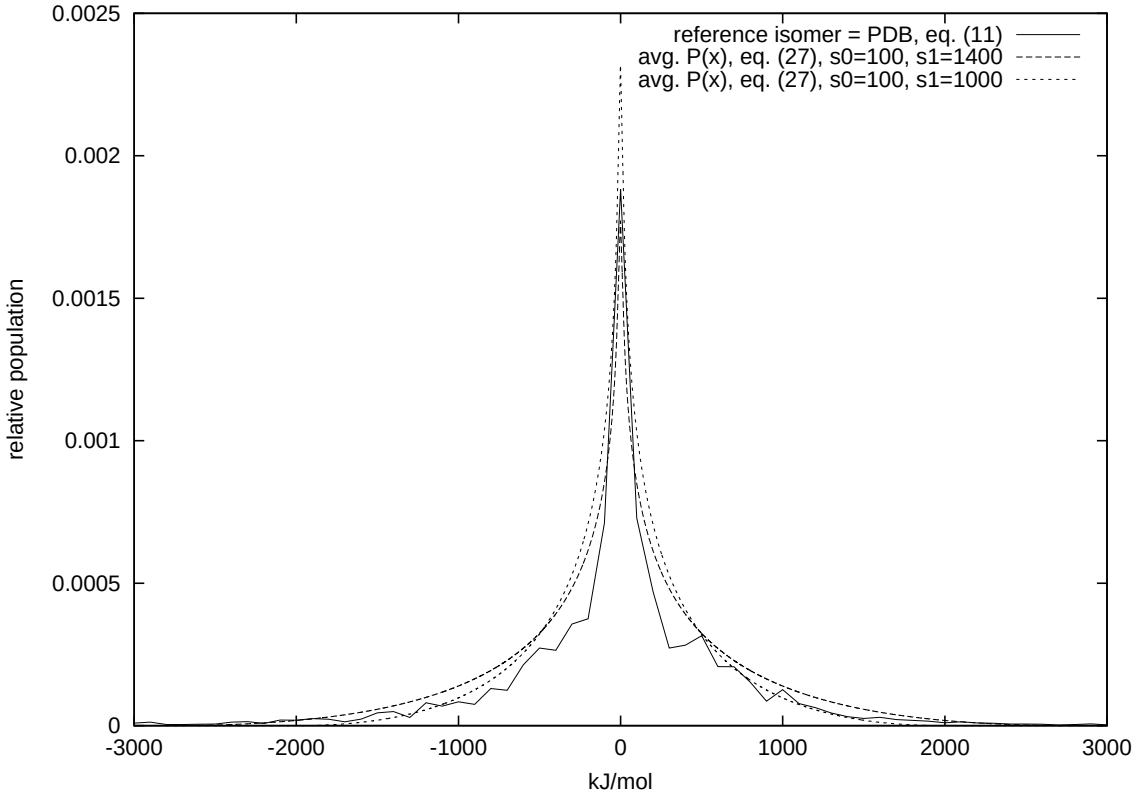
$$\bar{P}(X) = \begin{cases} L\left(\frac{-|X|}{4\sigma_1\sigma_0} + \frac{log(\frac{\sigma_1}{\sigma_0})}{2(\sigma_1-\sigma_0)}\right) & |X| < 2\sigma_0 \\ \frac{L}{\sigma_1-\sigma_0}(\frac{|X|}{4\sigma_1} - \frac{1}{2} + \frac{1}{2}log\frac{2\sigma_1}{|X|}) & |X| \in [2\sigma_0, 2\sigma_1] \\ 0 & |X| > 2\sigma_1 \end{cases} \tag{27}$$

which, surprisingly, *is almost identical* to the distribution obtained from Gaussians with the

same values of σ_0 e σ_1:



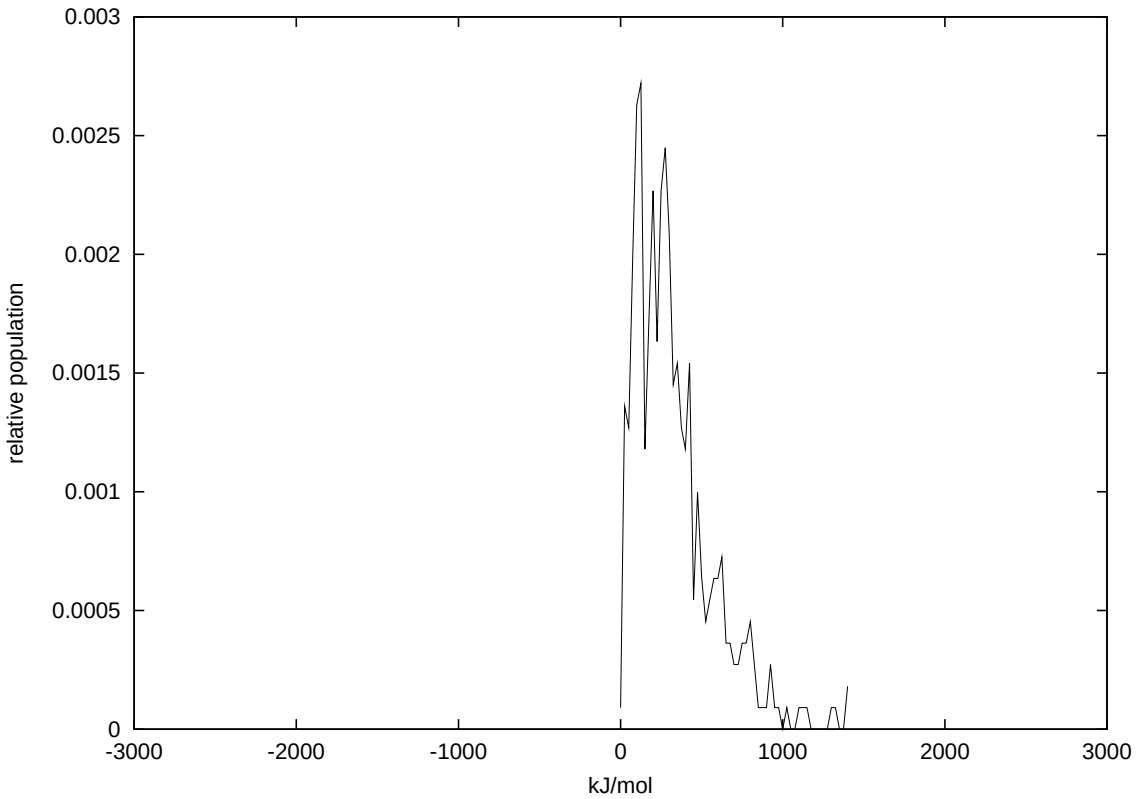
The comparison of the calculated function, for two choices of the σ range, with the observed "cleaned" distribution $p(X)$ is reported in the following figure: (409 proteins, 267 of which contain 1 to 2 histidines):



2.3.2 dependence on σ

The next approximation step is to consider a nonuniform distribution of σ, and to feed the right expression for $\rho(\sigma)$ in (23).

In the case illustrated in the accompanying paper, for example, (if σ is the standard deviation of the energy distribution of each protein) $\rho(\sigma)$ raises in the interval 10 to ~ 100, then decreases constantly from 100 to 1400:

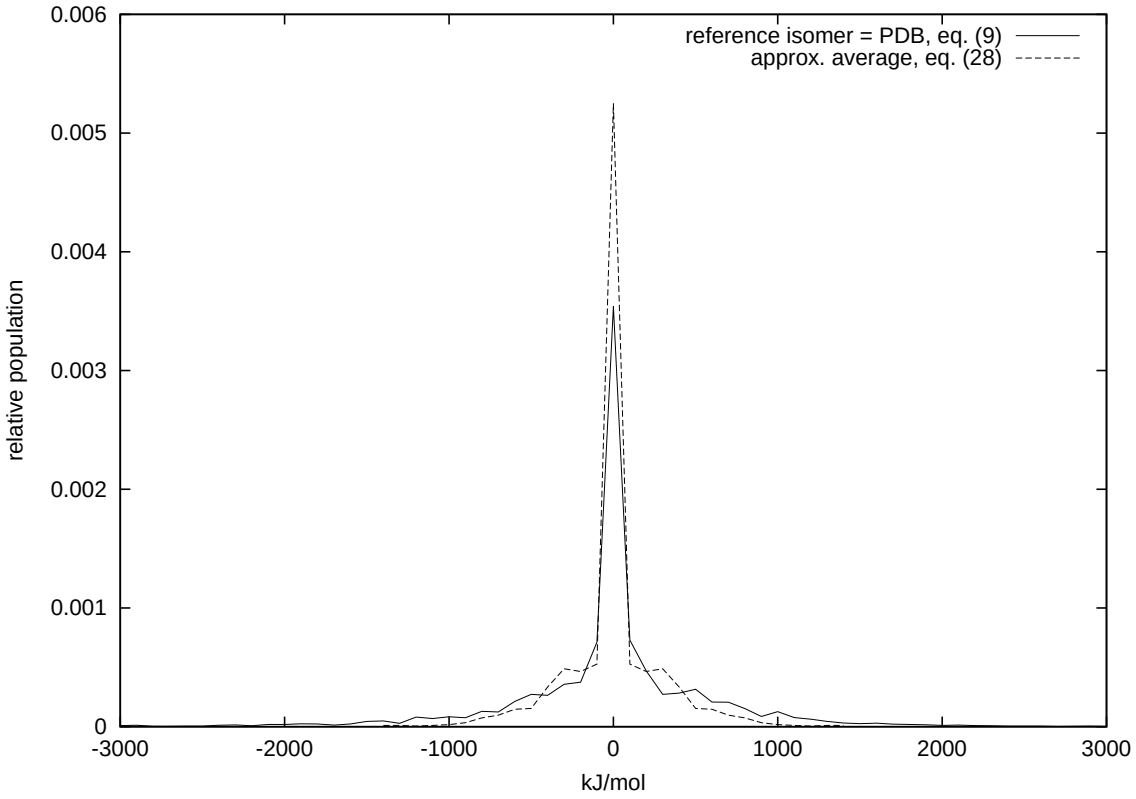


Note that the high fraction of proteins with low σ explains why the models just illustrated overestimate the distribution wings and underestimate the central peak.

A simple expression for $\bar{P}(X)$ can be written in the simple case presented in the previous section: if all proteins have only two isomers, whose energies are spaced by σ, the distribution will result (substituting (17) with $\sigma \equiv a - b$ into (23))

$$\bar{P}(X) = \frac{1}{4}\left[\rho(-X) + \rho(X)\right] + \frac{1}{2}\delta(X) \tag{28}$$

Using the observed value for $\rho(\sigma)$ one obtains



Here the central peak and the middle ($100-300$ kJ/mol) energy range steal population from the other regions, since continuous distributions have been replaced by three-pulse distributions of eq. (17). However, this is again a good approximation to the real $P(X)$.

3 Conclusions

In conclusion, **the average global distribution of energy differences is the sum of the correlation functions of each protein's distribution. These are even functions with a maximum in the origin that individually tend, with growing number of independent groups (histidines), to a Gaussian shape. The global distribution, if an equal weight is given to each protein, is still an even function with maximum in zero, however it has no Gaussian, or bell, shape, but a cusped or "witch's hat" shape, and this feature may be enhanced by a high proportion of distributions with low variance**.

References

[1] R. N. Bracewell, *The Fourier Transform and its Applications*, McGraw-Hill, 2000, ISBN 0-07-116043-4