ECOLE NORMALE SUPÉRIEURE

MASTER 2 : BIOLOGIE DES SYSTÈMES CELLULAIRES

Inference of WW domain Sequence-Function relationship using noisy measures.

Author:

Léo BLONDEL

Supervisors:

Aleksandra WALCZAK, LPT ENS

Thierry MORA, LPS ENS

Abstract

Protein's function is ultimately determined by its amino-acid sequence which in turn determine its structure. The general problem of the sequence-function relationship is very complicated and some case studies that focus on a well-defined function in a small portion of the sequence space can be informative. Here, we show that using a protein display assay coupled with high-throughput sequencing, it is possible to examine quantitatively the effect of thousands of mutations around a reference sequence. This approach rely solely on the sequences-function relationship and can provide valuable insights on the mechanisms of a protein without prior structural knowledge. The ability to extract valuable knowledge in the noisy context of biological experiments is made possible thanks to new inference techniques based on Information Theory. Using a simple model that considers each amino acid as independent we demonstrate the ability to infer an energy model for the binding of the hYap65 WW domain, which recapitulates to a certain extent the thermodynamics laws of a binding event. Finally, we extended our analysis by looking at the sequence-structure relationship of two other WW domains.

1 Introduction

Proteins are the key functional elements of living organisms. They are composed of a chained sequence of amino-acid. The physico-chemistry of those amino-acids in turn gives rise to the structure of the proteins as the chain folds. The structure then determines the function of the protein. Understanding the link between the sequence and the function of a protein has always been one of the main questions of biology. In order to understand the sequence-function relationship, the main tool that has been widely used is mutagenesis [1–3]. Indeed a single mutation can impact the catalytic activity, the structure, the stability, etc, of the protein allowing the study of the function of given amino-acids at given positions. Nevertheless understanding how a mutation impacts a protein usually requires specialized assays to measure such modifications and those are not scalable to high throughput approaches, preventing a more complete studies of proteins. As high-throughput sequencing becomes more easily accessible, new methods were invented to measure such properties on a large scale. One of those, called Phage Display, consist of displaying the protein of interest on the surface of the capsid of a phage Lambda T7, selecting the phage for a given function and finally sequencing before and after selection in order to indirectly measure the chosen activity [3–5].

The WW domain is a short protein domain, whose function is to recognize and bind to a specific peptide ligand. It is present in a lot of organisms and was named after a conserved common feature: two conserved Tryptophane (W) amino-acids at position 3 and 25. The Human Yes Associated Protein 65 (hYap65) contains two WW domain, and one of them was mutated and studied experimentaly by the Fields group and is the focus of this theoretical investigation [4, 5]. In 2010, Fowler *et al.* cloned a 45 amino acid portion of the first WW domain of hYap65, and 33 amino acids of that sequence were mutagenised to create 600 000 mutants containing all single and double mutations, and some of the triples. Then, those sequences were introduced into the T7 Lambda Phage genome,

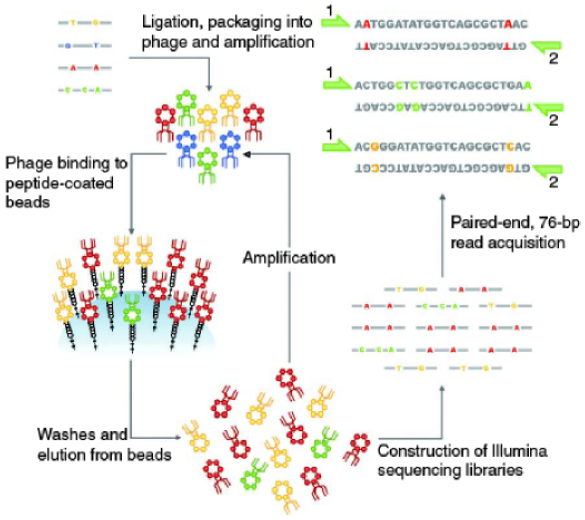

Figure 1: *Phage Display Procedure [4]*

fused with the capsid protein and amplified. Mixed with a proper sensor or ligand, this assay allows for the measurement of any activity the protein might have. In this study the authors were interested in it's binding affinity. The different phages were thus selected on magnetic beads coated with the hYap65 ligand, the beads were then extracted and subjected to washing cycles. The phages that stayed attached were then amplified on bacteria, purified, and again selected against the ligand. At round 0 (input mutant library), 3 and 6, parts of the phages were digested. Due to the technical limitations of Illumina paired-end sequencing at that time, only 76 bp (i.e. 25 of the mutated amino acid sequence) were amplified by PCR. Finally the resulting DNA was sent for high throughput sequencing (Figure 1) (For the complete procedure see Fowler *et al.* 2010) [4]. Using the number of read counts of each mutant from the sequencing they calculated the enrichment ratio and fitness scores of the 25 amino acids. They used statistical tools to study the relationship between the sequence and function of the WW domain. Here we propose a novel way of studying that relationship.

In 2006 and 2010, Kinney *et al.* developped a method to fit model parameters using Information Theory [6, 7]. Indeed, the main problem with the Phage Display experiment (as it is for many biological procedures) is the difficulty to estimate the error model, preventing any correct calculation of a likelihood estimator which depends on it. The classical way to overcome this limitation is to try to measure part of the noise and/or assume it follows a Gaussian distribution and use it to infer parameters. While the former is challenging, the later is usually easier. This approach proves to be correct in some cases but it often leads to a wrong parametrization of the model. Moreover, assuming a given error model returns only one set of parameters that is the most consistent one with the observation. Here, using the theoretical ideas developped by Kinney *et al.*, we extended information theory based the approach to protein binding energy modeling.

Our model here is a function that takes a protein amino acid sequence, a set of parameters θ and returns a binding energy E. Experimental data from the previous study of Fowler *et al.* 2010 are DNA sequence read counts, x_i, in a given run : z_i. A given set of parameters θ with a fixed error model will estimate the experimental data with a probability $p(\{z_i\} \mid \theta)$ also called likelihood. As the error model is unknown Kinney *et al.* used Bayes theory [7]. Assuming a prior distribution $p(\theta)$ on model parameters, one can write the posterior distribution on the model parameters:
$p(\theta \mid \{z_i\}) \propto p(\{z_i\} \mid \theta) \cdot p(\theta)$.

But in order to calculate $p(\{z_i\} \mid \theta)$ one still needs the error model which is not usually available, nor measurable. Kinney *et al.*, then averaged $p(\{z_i\} \mid \theta)$ over all possible error models and found that $p(\{z_i\} \mid \theta) = e^{N[I(z;x)-H(z)-\Delta]}$ where z is a measure and x an output of the model (in our case $x = E(\sigma_i;\theta)$, see *Materials and Methods: Error Model Averaged Likelihood*) [7]. This make the assumption that the error model is the same for all independent measurements in the experiment, and in Fowler *et al.* experiments it is very likely a valid assumption, as each phage will be subjected to the same manipulations, even if each round might have a different error model. Moreover, the following assumptions are made:

1) There exist a correct set of parameters θ that accurately describe the protein binding energy.
2) The likelihood of the data $\{z_i\}$ depends on θ only through the model prediction $\{E(\sigma_i;\theta)\}$:

$$p(\{z_i\} \mid \theta) = p(\{z_i\} \mid \{E(\sigma_i;\theta)\})$$

3) The measurement of each sequence σ_i is independent such that each sequence σ_i can be reduced to a single predictive real number $x = E(\sigma_i;\theta)$:

$$p(\{z_i\} \mid \{E(\sigma_i;\theta)\}) = \prod_i p(z_i \mid E(\sigma_i;\theta))$$

1.1 Information Theory

Mutual information (MI) between two random variables Z and X is defined as the amount of uncertainty (also called entropy) reduction about one variable gained by from measuring another random variable. Intuitively it can be seen as a measure of the quantity of information that can be known about one random variable by measuring another one. Given $\{z\}$ and $\{x\}$ two set of measures of Z and X, $I(z;x)$, can be written as such :

$$I(z;x) = \sum_{z,x} p(z,x) \cdot \log \frac{p(z,x)}{p(x)p(z)} \tag{1}$$

Specifically in our case the MI is written as :

$$I(z;x) = \sum_{j}^{Z} \sum_{i}^{N} p(x_i \mid z_j) \cdot p(z_j) \cdot \log \frac{p(x_i \mid z_j)}{p(x_i)} \tag{2}$$

1.2 Models

We started our modeling with the simplest energy model. It makes the assumption that the binding energy of the WW domain can be simplified as an additive model, i.e. each amino-acid is independent and participates additively with a given energy ε to the total energy E of a sequence σ, and is written as follow:

$$E(\sigma;\theta) = \sum_{p=1}^{P} \varepsilon_{i_p} \ , \ with \ \sigma = (i_1, i_2, ..., i_p) \tag{3}$$

$$\theta = \begin{bmatrix} \varepsilon_{1_1} & ... & \varepsilon_{1_p} \\ ... & ... & ... \\ \varepsilon_{i_1} & ... & \varepsilon_{i_p} \end{bmatrix}$$

$$i_p = \text{"Amino acid at position p"}$$
$$p = \text{"Position in sequence"}$$

Where σ is an amino acid sequence of length p, θ is a matrix of size $P * 20$, and $E_\theta(\sigma)$ the binding energy.

The second model we worked on was a more detailed energy model taking into account the relationship between amino acid (also called epistasis). It consists of two term, the single site energy which depends only on the amino acid at a given position (which can be seen as the interaction between that amino acid and itself) and an interaction term which depends on two amino acids at two different positions in the sequence. It is written as follows :

$$\begin{aligned} E_\theta(\sigma) &= H(\sigma) + J(\sigma) \\ E_\theta(\sigma) &= \sum_{p}^{P} \varepsilon_{a_p} + \sum_{p_1, p_2; p_1 > p_2} J_{p_2 b}^{p_1 a} \end{aligned} \tag{4}$$

$$\theta = \begin{bmatrix} \varepsilon_{1_1} & J_2^1 & ... & ... & ... & ... & J_{p_2 b}^1 \\ 0 & \varepsilon_{1_2} & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & ... & ... & ... \\ ... & ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & \varepsilon_{1_p} & ... & J_{p_2 b}^{p_1 a} \\ 0 & 0 & 0 & ... & 0 & ... & ... \\ 0 & 0 & 0 & ... & 0 & 0 & \varepsilon_{a_p} \end{bmatrix}$$

$$a, b = \text{"Amino Acid"}$$
$$p_1, p_2 = \text{"Position in sequence"}$$

Where, $H(\sigma)$ is the independent energy term, $J(\sigma)$ the interaction term, p_1 and p_2 the position in the sequence and a and b the amino acids at those positions. The practical difficulty with this model is that it is computationally very costly to simulate it with all Js (in our case 125 000). For this reason a small subset of non-zero Js need to be chosen beforehand.

Using a Simulated Annealing Monte Carlo simulation maximizing the MI, on the simple, independent-site model we were able to infer a set of parameters that describes the binding energy and follows Boltzmann binding characteristics over a certain energy range. We used a Monte Carlo algorithm that randomly moves one parameter, then it calculates the probability distribution of energy using the new parameter set θ. Next, the MI between the probability distribution of energy and the selection rounds is calculated. Intuitively, the MI tells us how informative the model is of the sequence composition in each round. Finally, the Monte Carlo algorithm, by making random moves, eventually maximizes the MI, i.e. it samples the region of parameter space that are the most informative of the measurement.

Finally, the more complex interacting amino acid model was explored but due to the time limitations of the internship results were not fully analyzed.

2 Result

2.1 Algorithm

The algorithm was written in pure python 2, but simulations were not feasible due to poor performance (6-8 s per Monte Carlo step which is equivalent to 100 days of simulations for 1 000 000 step). The code was first optimized to calculate each step with a minimum amount of calculations and fast access to stored data, mostly taking advantage of python dictionary hash tables, at the expense of memory consumption, and the average step time was then 3-4 s. Finally, the code was adapted for Cython compilation by defining a fixed type to all variables, as well as rewriting parts of the main loop function in C. Once compiled the average time was < 0.5 s per loop. For comparison, a less optimized C++ code was used and it's average step time was 1sec.

2.2 Phage Display Simulator - Independent Model

Due to the complexity of the simulation, we first decided to generate in silico data similar to those obtained from the sequencing. Indeed, using a simulator we can control the probability and error at each step, but most importantly, we know the energy model used to select the sequences. Using such an approach allowed us to control two things. First, that the algorithm was working, and secondly that it was inferring the right solution. Using the procedure described in the *Materials and Methods: Phage Display Simulator*, we simulated 6 selection rounds for a known energy model.

Four independent inference simulations were then run on sequences from that in silico dataset, as described in the *Materials and Methods: Monte Carlo Simulation*, and parameters were saved every 1000 steps. In order to reduce what is called the "heating" time of the Monte Carlo Simulation (the time necessary for the random sampling to find a correct set of parameters) the initial conditions were selected as the log-ratio of the frequency of each amino acid at a given position between the input sequences and round 6 of selection.

$$\theta_0 = \begin{bmatrix} \varepsilon_1^1 & \dots & \varepsilon_1^p \\ \dots & \dots & \dots \\ \varepsilon_i^1 & \dots & \varepsilon_i^p \end{bmatrix}, \text{ with } \varepsilon_{\mathrm{i}}^{\mathrm{p}} = \log\left(\frac{frequency_i^p(round_0)}{frequency_i^p(round_6)}\right)$$
$$i = \text{"Amino Acid"}$$
$$p = \text{"Position in sequence"}$$

Finally the MI was calculated as explained in *Materials and Methods: Mutual Information Calculation*.

During the inference runs, Information between the energy of the sequence and the selection round increased from 0.157 to 0.185 ± 0.0024 bits (mean over 4 runs). The simulation was then allowed to run another 200 000 steps and those 200 parameters samples were selected for analysis. The parameters were normalized as explained in *Materials and Methods:Normalization*

We first measured the IntraRun and InterRun variance of the inferred energy parameters (see *Materials and Methods: Intra and Inter run variance calculation*). Indeed our Monte Carlo Simulation samples the distribution of parameters and one of the common measure of sampling is the InterRun and IntraRun Variance. The InterRun variance measures the variances of the distribution

of the inferred energy parameters between the runs at each time point, thus measuring "how far" they are from each other. If it increases, the simulation parameters are further apart, while if it decreases, the simulation parameters are getting closer. It is calculated by summing the variance of the distribution of each parameters in the runs. The IntraRun variance measures the variance in the distribution of parameters from time t to time $t+n$. It measures the sampling variance as a function of time, and should only increase or stay constant. Indeed, once sampled entirely, the distribution variance is constant, while as long as the distribution is not fully sampled it is increasing. It is measured by summing the variance of the distribution from t to t+n of each parameter. Figure 2a shows that the InterRun variance has saturated quickly while the IntraRun variance still increases but has almost saturated too. Thus, the Monte Carlo sampling quickly finds a solution that is common to all runs but in each run, it progresses more slowly toward the full sampling of the parameter distribution. Ideally, the IntraRun and InterRun variance should converge to the same value (which correspond to the uncertainty on the parameters that were inferred), but it would require extensive computational time and the precision with which results are acquired is sufficient.

The main drawback of using the variance as a measure is that it aggregates all parameters together and thus, detailed analysis is not possible. Another way to look at the parameters is to scatter plot the distribution for each parameter of one run versus another run (Figure 2b), and one run versus the known correct parameters (Figure 2c). This way, one can see the distribution of each parameter along with the precision with which they are inferred. Here, the InterRun scatter plot shows that all parameters have converged toward a common solution. Finally, when the distribution is plotted versus the known value we can see that most of the parameters are aligned but some in the extreme values lose a little bit of precision. Here, when looking at the precision with which we are able to infer the parameters, after 200 000 steps of simulation we are able to correlate our results between runs with a Pearson correlation coefficient of 0.997 ± 0.001. Moreover, each run recover the input energy model parameters with a Pearson Correlation coefficient of $0.988 \pm 7 \cdot 10^{-5}$ and with a correlation of 0.992 when averaging the 4 runs together.

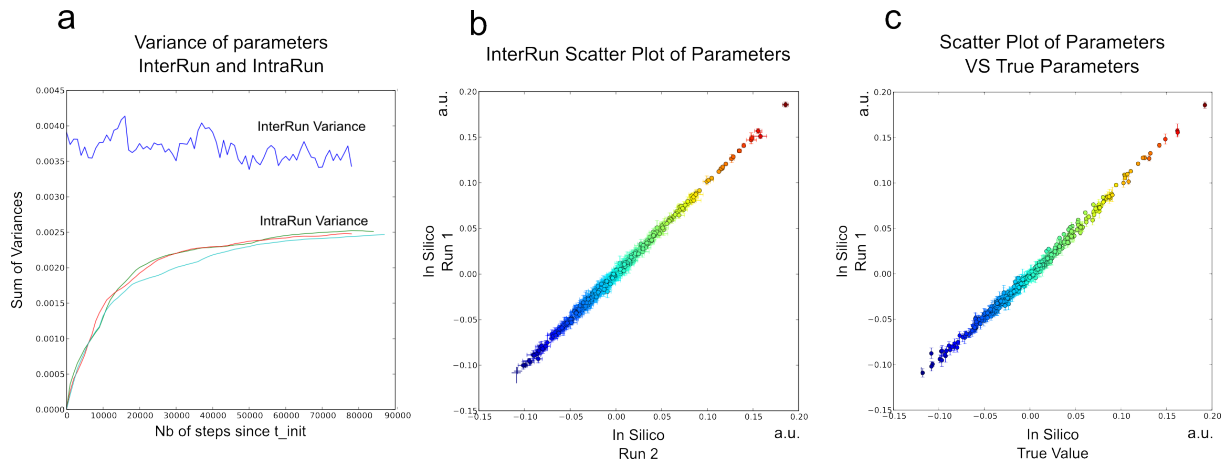

Figure 2: *In Silico Experiment : a) Inter and IntraRun variances of 3 simulations, b) Scatter plot of two independant runs, c) Scatter plot of an independant run versus the known parameters values*

2.3 hYap65 Independent Model fitting

The next step was to see whether our simulation could find the parameters of an unknown energy model with real protein data. To be able to run our Monte Carlo Simulation we needed real Phage Display data, thus we recovered sequencing data from the Fowler *et al.* [4] experiment as well as the Araya *et al.* [5] experiment. Those data were processed according to the procedure described in *Materials and Methods:Sequence assembly and quality filtration* in order to extract the different counts of the different mutant sequences at each selection round. Two sets of four Monte Carlo Simulations were run for 10^6 steps to draw independent measurement and parameters were sampled every 1000 steps. The inverse of the temperature β was increased exponentially until 10^6. In order to be comparable across time and runs, they were normalized as explained in *Materials and Methods:Normalization*.

As we previously observed with in Silico Data, the MI between the energy of the sequence and the selection round went up from 0.118 to 0.151 ± 0.0038 bits (mean over 8 runs) and then plateaued (Figure 3a). Moreover, the different runs quickly converge toward the same solution, as shown by the InterRun variance which quickly saturates around 0.055 (Figure 3b) As for the IntraRun variance it slowly increases but does not saturates (Figure 3c), which shows that the simulation has not sampled the whole parameters distribution. When looking at the scattering of the parameter in between runs, the Pearson correlation coefficient is good with a mean value of 0.987 ± 0.005 (Figure 3e-f). When pooling together the distributions of the first set of 4 simulations

and plotting it versus the other 4 of the second set, the scattering shows a high correlation with a Pearson correlation coefficient of 0.9978 ± 0.0001 (mean over all possible two sets of 4) (Figure 3g).

The fact that the IntraRun variance has not saturated shows that more simulation time is needed to sample the complete parameter distribution in a run, but because our runs are independent and, as shown by the InterRun variance, sample the same parameter space we can merge their result. When plotting the variance of the eights runs combined we can see that it saturates quickly around the same value as the InterRun variance (Figure 3d).

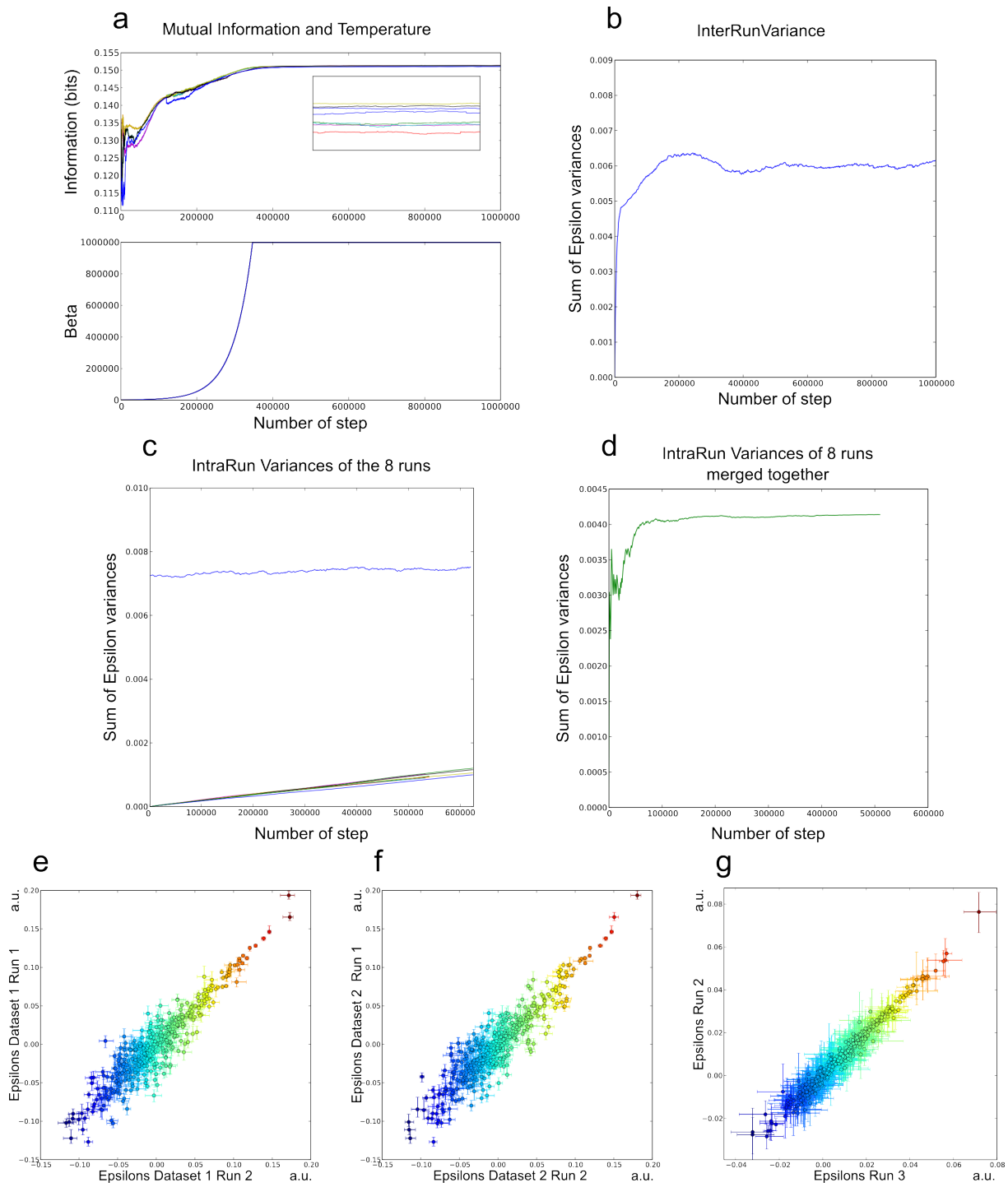

Figure 3: *Results of the Monte Carlo Simulation on hYap65 data a) MI and Inverse of temperature* β *of the 3 runs. Information saturates as the model samples the parameters space. b) InterRun Variance of the 3 simulations. After a quick increase when the temperature is high, the variance then stabilizes itself and decreases to a plateau at 0.06. c) IntraRun Variance of each simulation. Each run samples the space around it slowly as can be seen by the 3 curves. At the top is the InterRun variance which is the theoretical maximum that each run can reach. d) IntraRun Variance of mixed runs. For each run, parameters were pooled and shuffled into one distribution. The variance quickly saturates showing that when averaged multiple run quickly sampled the distribution space. e,f)Scatter Plot of the parameters distributions for each run versus another run. They globally find the same solution even if some parameters are less constrained than others. e:*$r^2 = 0.987$*,f:*$r^2 = 0.984$*. g)Scatter Plot of the parameters distributions of the two datasets.* $r^2 = 0.995$

2.4 Boltzmann Distribution

The goal of our model is to infer the binding energy of hYap65 to its ligand. The main theory that describes the thermodynamics of molecular micro-states was proposed by Boltzmann for gas molecules but has been generalized to a lot of systems. As explained in *Materials and Methods: Boltzmann Probability*, the probability of binding can be written as follows:

$$p_{binding}(\sigma) = \frac{1}{1+e^{-\beta\Delta E}} \ , \ \Delta E = E_{unboud} - E_{bound} \tag{5}$$

An important point of our simulation is that the inferred energy has to fit correctly the Boltzmann law. With the inferred energy matrix of the WW domain, we can calculate for each protein σ the energy term ΔE. Moreover, we know the frequency of each protein at round 0, 3 and 6 that is given by the number of reads of σ, normalized by the total number of reads for each round. Thus, we can plot the probability distribution of binding as a function of the sequence energy as shown in Figure 4a. In order to avoid discretization issues due to the binning of the continuous distribution, it was smoothed using a convolution with a Gaussian of variance 0.01 (see *Material and Methods: Mutual Information Calculation*) and is shown on Figure 4b. As expected by the normalization, the lower the energy, the higher the probability to be selected. But the absolute probability, while informative, can be normalized by a prior. Here we can normalized round 3 and 6 by the probability at round 0. As explained in *Materials and Methods:Boltzmann Probability*, the log ratio of the probability at round k in the case of high energy can be approximated as:

$$\log\left(\frac{p^{bound}_{round_k}(\sigma)}{p^{bound}_{round_0}(\sigma)}\right) = -k\beta\Delta E \tag{6}$$

This can be visualized in Figure 4c, where the log ratio of the smoothed probability at round 3 and 6 over the smoothed probability at round 0 are plotted against the energy. At very low and high ΔE there is not enough information in the data (for example at high ΔE the protein will very likely not fold and thus will not be seen at round 3 and 6) so that the precision decreases, but as predicted, at intermediate ΔE (high E_{bound} as $\Delta E = E_{unboud} - E_{bound}$) the log-ratio follows a near linear law, allowing us to check whether:

$$\frac{\log\left(\frac{p^{bound}_{round_3}(\sigma)}{p^{bound}_{round_0}(\sigma)}\right)}{k_3} = \frac{\log\left(\frac{p^{bound}_{round_6}(\sigma)}{p^{bound}_{round_0}(\sigma)}\right)}{k_6} = -\beta\Delta E \ , \quad \text{with}\, \mathrm{k}_3 = 3\, \text{and}\, \mathrm{k}_6 = 6$$

When we fit the curve, we measure that in order for round 3 to be equal to round 6, the factor is close to but slightly deviates from k. Indeed we find that $k_3 = 3.71 \pm 0.052$ and $k_6 = 5.17 \pm 0.038$ (mean over 8 runs). This multiplicative factor of 1.4 ± 0.001, shows that our predictions are not perfect. But this might be explained by two things. First, our model considers the energy as a simple additive law, which is an oversimplification. The second one is the fact that we lack the intermediate rounds 1, 2, 4 and 5 and thus our algorithm cannot completely infer the energy terms. Nevertheless, our approach still gives excellent result knowing that we are inferring the energy on highly indirect energy measurement, noisy rounds of selection/amplification, and finally, sequencing technology that do not perform precise measurement.

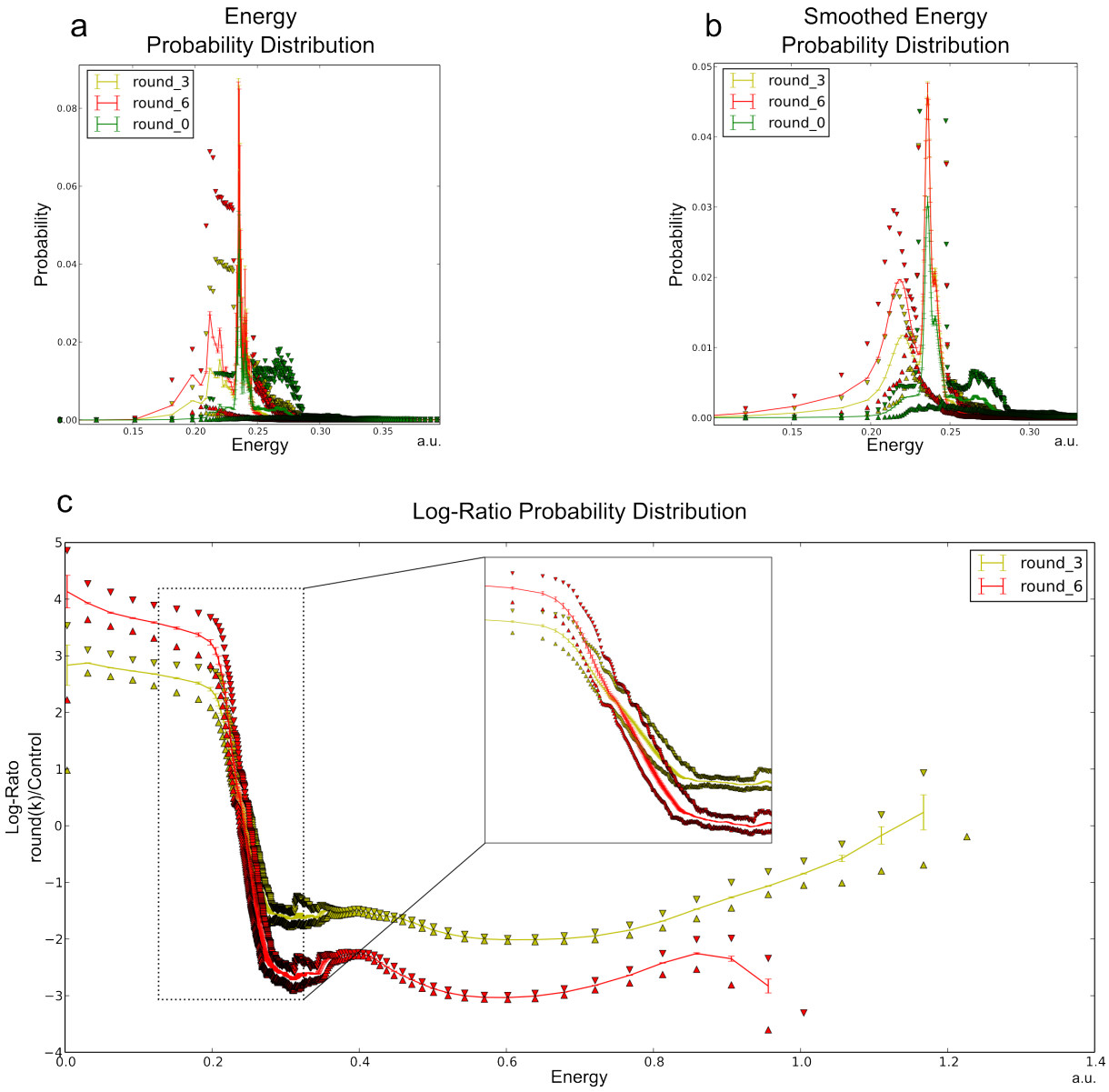

Figure 4: *Boltzmann energy model. The plain line represents the mean of the distribution, one standard deviation is plotted as the error bar and the min and max are shown as triangles a) Raw energy probability distribution calculated using inferred parameters. At round 0, the distribution has one peak which correspond to the Wild Type hYap65, and a long tail at high energy. At round 3 and 6, the tail almost disapears and a peak appears at low energy. b) Same as a, except that the distribution was smoothed by convoluting it with a Gaussian distribution of variance 10. c) Log Ratio of Probability distribution at round 3 and 6 divided by round 0. The linear part of the curve is zoomed in. At very high and low energy the precision diminishes as the data is not well sampled in those areaa.*

2.5 Energy unit fitting

The main problem with our inference algorithm is that MI is non parametric, and thus ΔE is unitless and can be inferred up to any multiplicative constant. But, using the log ratio of the energy

probability distribution, we can write:

$$\log\left(\frac{p^{bound}_{round_k}(\sigma)}{p^{bound}_{round_0}(\sigma)}\right) = -k\beta(\Delta E * const) \tag{7}$$

Moreover we can choose to express $\beta = \frac{1}{k_b T}$ in the units of our choice as $T = 300K$ and the unit is given by the Boltzmann constant k_b. As is common in biology, we choose to express our energy in Kilo Calorie per mole. If we divide the linear fit of the log-ratio distribution for the round 3 and 6 by $(-k\beta)$ and then average over the rounds and runs we find that the multiplicative constant is: 10.13 ± 1.62. Results of the energy matrix expressed in kcal/mol are shown in Figure 5 where the Wild Type sequence energy terms are squared in yellow.

When looking at the energy matrix, one can see that generally, the loops allow for more variety in amino acids compared to the beta strands. As was previously reported [4, 5], the WT protein is not the best ones when looking solely at the binding function. Indeed the energy terms of the WT at each position are not always the best one. The 18 values at 2.40 kCal/mol are amino acids that were counter selected so early that no information was carried to round 3 and 6 and thus their energy contribution cannot be trusted as it is the result of the pseudo-count used to avoid divisions by 0.

Finally, within our model, calculating the relative probability of each amino acid at each position would require to compute the energy for the 20^{25} possible sequences. Thus, we decided to approximate the relative probabilities by multiplying each energy contribution ε by the number of position in sequence (25) such that : $p(\varepsilon_i^p) = \frac{1}{1+e^{-\beta\varepsilon_i^p * 25}}$. The Logo was then created using those probabilities and is displayed on Figure 5.

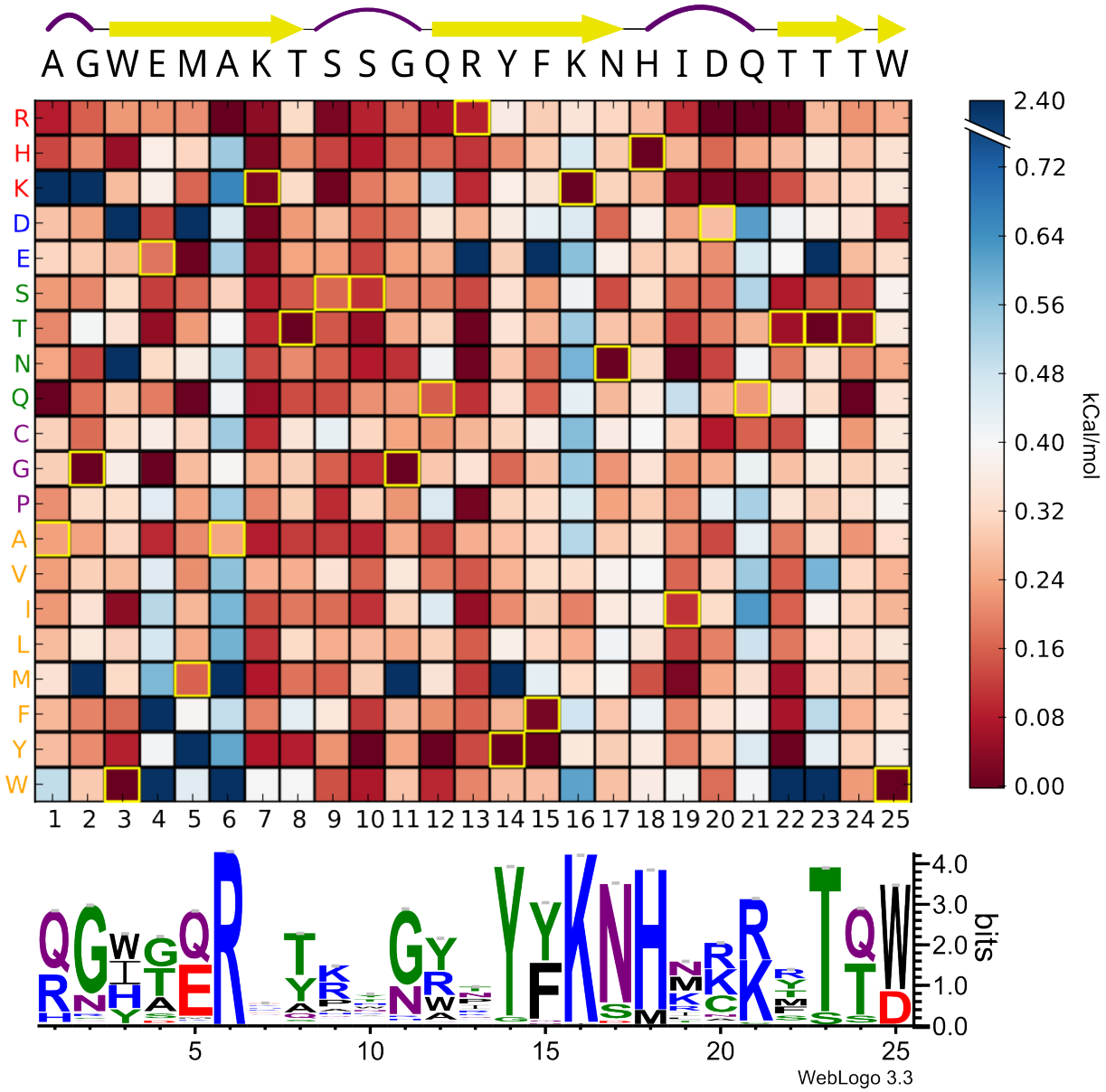

Figure 5: *Energy Matrix in kcal/mol of the energy participation of each amino acid to the total binding energy. The yellow square show the wild type hYap65 sequence. The logo was computed using WebLogo 3.3.*

2.6 Protein Analysis

Having inferred our energy matrix, we wanted to see if we could use it to match previous structural analysis of the WW domain. Most WW domains (but not all) have a W at position 3 and 25. Here, the energy matrix (Figure 5) shows that the two Ws are of low energy, and thus very stabilizing. But we also find that the Ws at position 3 and 25 are not exclusive amino acids for the binding of the peptide. Indeed, it was previously reported [8] that the molecule looses thermodynamic stability

and binding affinity on mutation of the 25th tryptophan to a phenylalanine but could still fold. It is completely unfolded when the 3rd tryptophan is mutated to a phenylalanine but can still bind by folding itself on the ligand. It is interesting to see here that our simulation predicts that the mutation W3F is of lower energy than the mutation W25F [8].

In order to understand better the relationship between the position and the variability in amino acid, we looked at the distribution of variances of each amino acid energy by position, as shown in Figure 6. We can clearly see that some amino acids are highly constrained, like the Q at position 1, W at position 3, E or Q at position 5, K at position 16, with a low energy contribution and a variance near 0. While most of the others have a high energy a higher variance. Moreover, each position possesses it's unique amino acid signature, with positions that allows one, two or three amino acids of low energy and the others are high, and positions where all amino acid are of high energy. This is linked to the level of constrains imposed on that position to maintain it's function.

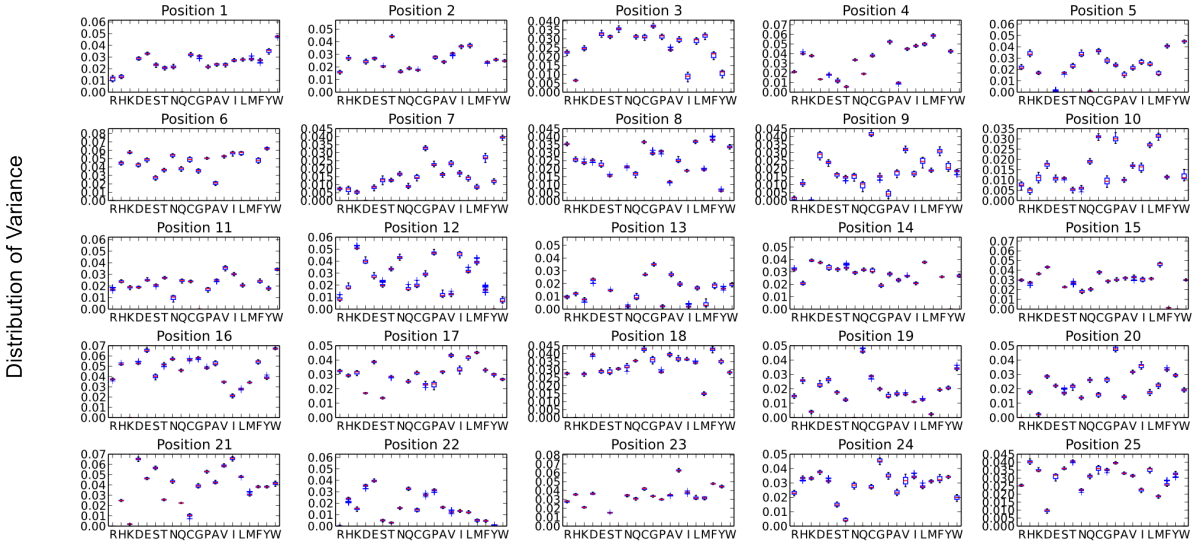

Figure 6: *Boxplot of the energy distribution of each parameter. Most of the parameters have a medium energy, with a higher variance. Some of them are really constrained, with a low variance and low energy. For example, the W in position 3 is at 0 thus it doesn't appear in the boxplot.*

2.7 Structural analysis

In order to visualize the inferred energy in the 3D context of the protein, we decided to show the structure colored according to the contribution in energy of each position. Each amino acid is colored from red to white, red being the lowest energy and white the highest. Even if, our model

inferred energy using mutants generated from the WT hYap65 sequence, the energy terms should apply to WW domains that are close to the WT but that were not in the input library. While this approach is questionable and should require a more thorough analysis, it still brings interesting results and helps to comprehend the usefulness of the modeling approach.

Using Pfam protein alignment and the PDBe Structure similarity tool [9] we chose a structurally related WW domain (from the protein NEDD4) and a well studied but unrelated WW domain (from the protein APBB1) that had NMR structures. We colored each position for the three proteins and the results are shown in Figure 7.

First, when looking at hYap65 (Figure 7a), we can see that, as previously reported [4], the loop parts of the protein are the less constrained with a high contribution in energy. It is also striking that the amino acid that contributes the most in the binding energy of the protein is the Threonin (T) at position 22 while the energy prediction (Figure 7) shows that an Arginine (R) or a Tyrosine (Y) would decrease the binding energy. It is understandable why a Y would be a better match as the ligand is mostly hydrophobic, but the T might bend the ligand in order for the W at position 25 to be in contact with the Proline (P). Finally, it is interesting to see how the energy is concentrated in the core of the protein, especially on the three beta sheets.

Next we looked at the human APBB1 protein WW domain (Figure 7c-d). When we colored the positions, it showed interesting properties. Indeed, the W in position 3 is located near the surface of the protein. It is normally needed to maintain a hydrophobic core inside the protein. Here we can see that it is stabilized by two other hydrophobic amino acids P and H, two of them with a low energy contribution. Moreover, the G is the most exterior amino acid and probably stabilizes the structure by counterbalancing the hydrophobicity of the three amino acids.

Finally we looked at the E3 ubiquitin-protein ligase NEDD4 3rd WW domain (Figure 7b). The same properties as for the hYap65 WW domain emerged, the core was low in energy contribution, with the same important T at position 22.

Even if those results need to be taken with caution, as the inferred energies cannot be taken as a precise measurement, it is still very interesting that patterns of stabilizing elements appear in the structure. Moreover, it seems clear that the energy encompass two properties, the stabilization of the structure and the binding affinity. Indeed, using that phage display experiment it is not possible with our methodology to separate those two energies. For example, the W at position 3 never comes

into contact with the ligand but participates as much as the W at position 25 which plays a major role in the binding affinity [5].

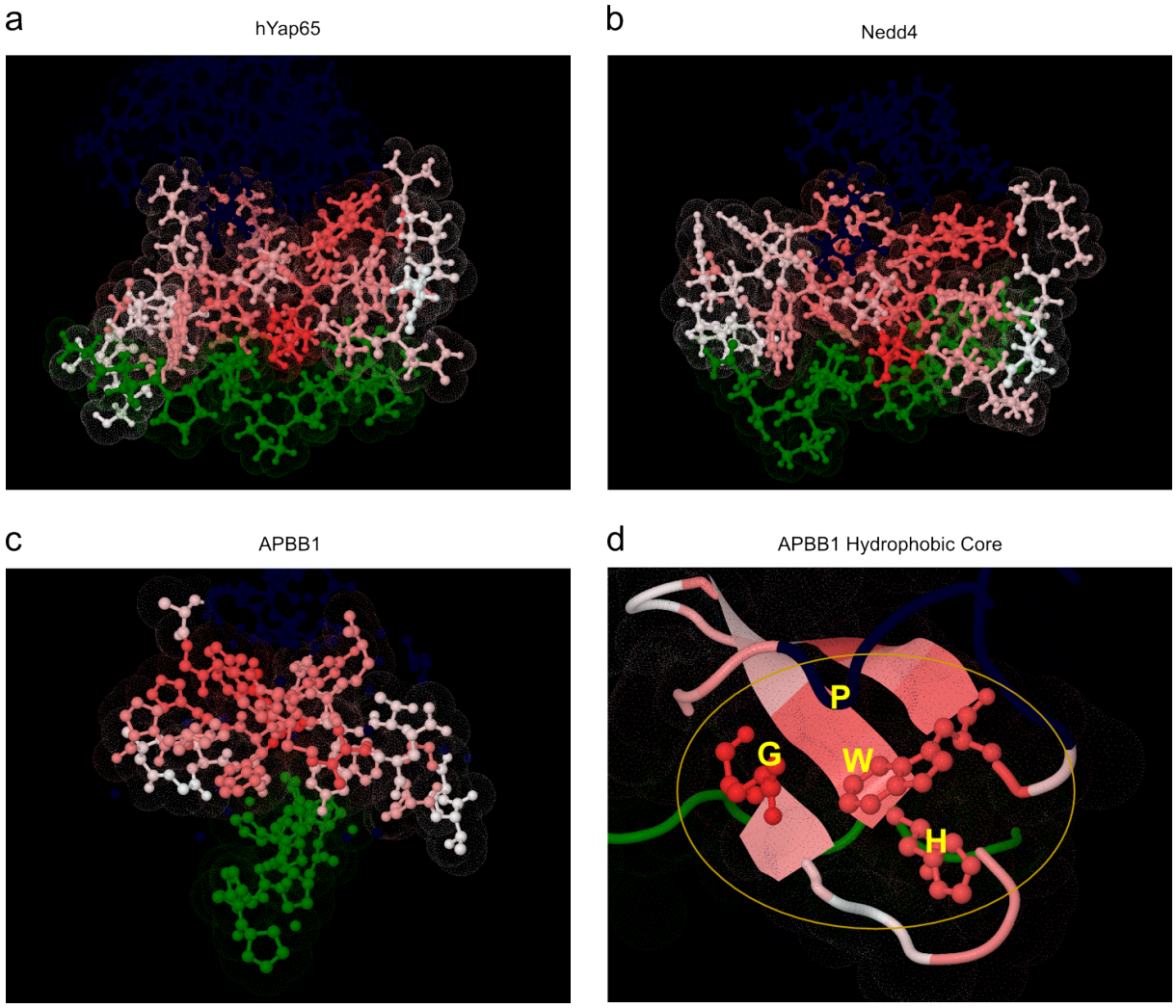

Figure 7: *NMR structures of WW domain. All structures were colored with the inferred energy, red being the lowest energy while white is high energy. The ligand is pictured in green and the rest of the amino acid of the WW domains are in dark blue. a) hYap65 3rd WW domain. b) NEDD4 WW domain c) ABPP1 WW domain. d) ABPP1 W at position 3. Here the 3 amino-acid are hydrophobic but in contact with water, thus all 3 contributes strongly to the stability of the structure.*

2.8 Comparison to literature

Interestingly, we can also predict some of the behavior of the mutants that were described in Araya *et al.* 2012 [5] with our independent model. As shown in Table 1, L30I, Q35K and D34T have a higher binding probability due to a lower energy, and I33R, T36R has a similar binding probability. Nevertheless, those mutations are supposed to be “stabilizing” mutations, and our inference infer

the energy both on stability of the protein and specificity/affinity energy of the binding. Thus this explains why our prediction does not completely match the measurements of Araya *et al.* 2012 [5] (Q35K has a higher probability than D34T, but has a lower Tm and thus is less sable). Moreover, it is not possible to discriminates between stabilizing and binding energy in our results. To do so, we would need, for example, to mutate not only the protein but also the ligand in order to compare for the same protein and the same ligand different inferred energies.

Mutation	Sequence	Binding Probability	Araya 2012 T_m
WT	AGWEMAKTSSGQRYFLNHIDQTTTW	50,91	43.0 ± 0.5 %
L30I	AGWEMAKTSSGQRYFINHIDQTTTW	53,31	54.8 ± 0.5%
I33R	AGWEMAKTSSGQRYFLNHRDQTTTW	49,14	41.7 ± 0.5 %
D34T	AGWEMAKTSSGQRYFLNHITQTTTW	54,60	49.6 ± 0.4%
Q35K	AGWEMAKTSSGQRYFLNHIDKTTTW	59,40	46.6 ± 0.6%
T36R	AGWEMAKTSSGQRYFLNHIDQRTTW	51,73	38.4 ± 0.6 %

Table 1: *Binding probability of hYap65 Mutant calculated using Boltzmann probability law (Equation 5) and the inferred parameters. The Araya 2012 T_m are the melting temperature of listed variants as measured by Araya et al. [5]*

2.9 Interacting Amino Acid Energy Model

As a next step, we consider a model with increased complexity where we allow amino acid interaction as described by Equation 4. The model was implemented in the inference algorithm and the code was modified so that the computation time follows a law in O(n) where n is the number of Js selected, meaning that the time increased linearly with the number of Js selected for the fit.

Moreover, due to the sparsity of the J matrix, parameters are normalized automatically due to the fixed 0 value for most of the Js. Due to the internship deadline, the J selection prior to parameter fitting using the Monte Carlo was not finished, as is the analysis of the our in silico interaction model.

3 Discussion

Here we show that the methodology used by Kinney *et al.* [6, 7] to infer DNA-Protein binding energy can be applied to the modeling of the more complex, and knowingly non linear, protein-protein binding energies. The approach we used to compute the energy model of the WW domain

can be applied to a lot of different proteins. Indeed, no prior knowledge of the structure is needed. Once the functional phage display assay, or any other high throughput measurement method is performed, the different energy terms of the protein can be inferred, and the structure stability, functional mutant, or any function can be studied at the amino acid level without having to perform expensive and time consuming biochemical experiments. Nevertheless, a lot of improvements remain to be done. Indeed our simple additive model does not recapitulate all energetic interactions that happen in the protein. For example, we cannot discriminate between energetic contributions that stabilize the protein and energetic contributions that come from binding the ligand. This would require measuring additional information, for example performing the same experiments but under different temperatures. Then it might be possible to infer the stability energy terms and at the same time the binding energy terms.

Using MI to infer the energy of the protein can be done without knowing the error model, which in this case is essential as it is very hard to measure. With this approach, we sample at the same time the error model distribution and the parameters distribution. However, due to the non parametric property of MI, the parameters are necessarily unit-less. We overcame this problem by using the fact that our system is bimodal and thus can be modeled using the simple binding Equation 5. This equation can be approximated for its linear part which allowed us to restore units in the predicted energy using the Boltzmann constant.

Using our inferred energy we looked at the structure of WW domains to see whether our inferred energy would show patterns in the amino acid arrangement. Indeed we saw that our energy seemed to correlate with groups of amino acid in contact. Moreover, the core of the domain had a lower energy, consistent with the globular property of the WW domain. The loops showed a higher variance, also consistent with the increased flexibility of that part of protein. Also, looking at variants of our protein for which the structure had been resolved, we noticed that the energy could also highlight some of the structural properties. This result probably indicates that our inferred energy is composed of two part, one part that describes the stability and another part which accounts for the affinity. Indeed, those variants are not supposed to bind the same ligand as the hYap65 but still harbor some energetic patterns like the W at the core of the protein that stabilizes the globular structure. As a next step, using our energy to predict structure folding will be considered. Indeed the WW domain is a case study of protein folding and it would be interesting to see whether we

can predict the structure of the hYap65 and other WW domains and at which level of precision.

Another interesting property of the inferred energy matrix is the fact that it shows that the two Ws at position 3 and 25 are not necessary for the protein to bind to its ligand. Almost all known WW domains present naturally in a variety of organisms harbor those two Ws. This raises the question of some unknown properties about those Ws that plays a role in vivo. A hypothesis would be that even if the protein can bind to its ligand without one or two of the Ws, it might not fold properly [8], and thus be eliminated in the cell through the proteasome mediated pathway. Another hypothesis would be that the ligand usually does not present itself alone and is part of more complex structure, thus, the in silico experiments may not mimic entirely the binding event in vivo and the whole target protein should be used in the Phage Display assay.

Furthermore, it is interesting to see that the WT protein is not the best protein, indeed, other mutants have a higher binding affinity. From an evolutionary perspective, that seems contradictory, as evolution (if applied on binding energy) would have, very likely but not necessarily, selected mutants with a lower energy. This probably shows that some counter selective pressure keeps the hYap65 WW domain as it is. One might guess that the capability to bind to slightly different ligands might be one of those pressures. Indeed, the selection experiment was performed on a unique small peptide. In vivo, the hYap65 protein is known to bind to multiples proteins like Smad7 [10], or Runx2 [11]. Also, if the binding of the two proteins is too strong, regulation by an unbinding event might be inhibited.

Nevertheless, the simple model does not predict all properties, like epistasis, and thus a more complete model should be studied. To do so, the relationship between couples of amino acids in the protein will be studied using the Araya *et al.* [5] data. The 500 highest and lowest epistasis term will then be selected as J and simulations will be run.

Moreover, inferred energy could be tested against in silico folding predictions. Indeed, using Rosetta's framework, one can predict the folding and binding of a protein given its sequence and its ligand.

Finally, our simple model still reflects a lot of interesting properties that could easily be tested in a lab. Indeed, it would be interesting to measure the binding energy of some of the mutants, as well as their thermodynamic stability to highlight some of the properties that appear in the energy matrix.

4 Materials and Methods

4.1 Sequence assembly and quality filtration

Next Generation Sequencing data were obtained from the NCBI SRA website. Two dataset were used, SRA020603 [4] and SRA058752 [5].

FastQ files were extracted for each experiment. Then paired ended sequences were realigned using an alignment tool that scored 1 for aligned bases and -1 for non aligned bases and did not allow for gaps. Aligned sequences were then fused into one sequence and the overlapping bases quality scores were multiplied. If there was a contradiction between two bases, the one with the higher score was selected, if both scores were equal then the sequence was discarded. Then, sequences with a Phred score superior or equal to 20 were kept. Finally, DNA sequences were translated into protein sequences using the open reading frame that maximized the alignment score with the wild-type sequence.

4.2 Error Model Averaged Likelihood

Let us assume an experiment Z that measures a given phenomena X. Each measurement z_i will describe with a certain probability the correct value x_i. That probability is called the error model $E(z_i \mid x_i)$. In our case, $\{z_i\}$ is the selection round ($\{z_i\} = \{0,3,6\}$), x_i the binding energy of a protein and N = 714701 protein sequences. Knowing the error model one can write :

$$p(\{z_i\} \mid \theta) = \prod_i E(z_i \mid x_i) = \prod_{z,x} E(z \mid x)^{c_z x} \tag{8}$$

where $c_z x$ is the number of bins assigned to bin Z and X.

But, firstly, next generation sequencing error models are unknown, and secondly, even if the technical error model was known, the highly indirect measurement of binding energy through rounds of selection makes it almost impossible to know the error model.

As previously described [7], one can then average over all Error models given an acceptable

prior, which in our case is a uniform probability over all Error models:

$$p(\{z_i\} \mid \theta) = \int dE \cdot p(E) \cdot \prod_{z,x} E(z \mid x)^{c_{z,x}} \tag{9}$$

$$p(\{z_i\} \mid \theta) = e^{N[I(z;x)-H(z)-\Delta]} \tag{10}$$

$$\ln(p(\{z_i\} \mid \theta)) = N[I(z;x) - H(z) - \Delta] \tag{11}$$

Here, $I(z;x)$ is the empirical MI between the round and the binding energy calculated using observations and model predictions, $H(z)$ is the empirical entropy of $\{z_i\}$ and does not depend on the parameters, and Δ is a constant nonnegative correction term, for finite data and the error model prior $p(E)$, which tends to zero when N is large. Here, only $I(z;x)$ depends on the model parameters through x. The other terms depend only on the data. Thus, the best model will be the one that maximizes the MI given the experimental result, regardless of how those results were acquired, ie, regardless of the error model. Indeed, both the parameters and the error models will be sampled to find the combination that best fit the data.

4.3 Mutual Information Calculation

The energy for each sequence is calculated and a histogram of 1000 equipopulated bins is created using the energy distribution for each round, and one for all sequences of all rounds pooled together. Each bin is then filled with the frequency of reads. Each histogram is then smoothed using a Gaussian kernel of variance $\sigma = 0.01 * number\ of\ bins$, in bin units, and normalized by the sum of the bins values. The histograms now represent the discretized energy probability distribution and can be used to calculate the MI with Equation 2. Indeed, the histogram for each round is $p(x_i \mid z_j)$, the histogram regardless of the round is $p(x_i)$ and finally $p(z_j) = \frac{\text{Number of read in round j}}{\text{Total number of reads}}$.

4.4 Phage Display Simulator

We generated 600 000 sequences by mutating 1-3 amino acids from the reference sequence corresponding to the Wild Type sequence of hYap65. Each sequence was then amplified randomly until 10 millions unique "phage" were counted. Next each "phage" was selected for binding using a simple binding probability law :

$$p_{bound}(\sigma) = \frac{1}{1+e^{-E(\sigma)}} + \eta \tag{12}$$

where η is a Gaussian noise of variance 0.01. All the unbounded phage were discarded and the bounded ones were "sequenced", then re-amplified until 10 millions "phages" were counted. Sequencing was done with an error probability of $\frac{1\,\text{base}}{10^6\,\text{bases}}$.

4.5 Monte Carlo Simulation

Monte Carlo Simulations allow sampling of the probability distribution of parameters, making random moves in the parameter space, then assessing if the moves should be kept according to an acceptance function. Here, we defined our acceptance function (also called Metropolis acceptance rule) :

$$p(acceptance) = e^{\beta(I(\{z_i\}|\{x_i\}_{\theta'}) - I(\{z_i\}|\{x_i\}_{\theta}))} \tag{13}$$

Where θ' is the new set of parameters generated randomly, $\{x_i\}'_\theta$ the new model output calculated with θ' and $I(\{z_i\} \mid \{x_i\}_{\theta'})$ the MI between the new model prediction and the observations. In our case we used a modified version of Monte Carlo called simulated annealing that was created by Metropolis and Hasting [12]. The main concept is that we simulate a heat-cool process by driving the temperature term β from a low value (high temperature) to a big value (low temperature). This tweaks the amount of "bad" moves the algorithm accepts as it amplify the influence of the subtraction of Equation 13. The intuition for it comes from the formation of crystal structures. A crystal is the lowest energy state a molecule can be in, if they are heated then cooled brutally they will be fixed in a non optimal state of energy, but if the cooling is slow enough, molecules will have time to sample the space to form a crystal structure that minimizes the energy. The same is true for our parameters, if heated and then allowed to cool slowly, the chance of finding the global maximum (here we maximize information) is higher. For a more detailed explanation of the algorithm see Figure 8.

The algorithm was coded in Python, and modified to be compiled using Cython for performance improvement. Different libraries were used to perform the simulation and analysis. Numpy, Scipy [13], Matplotlib [14], Cython [15]

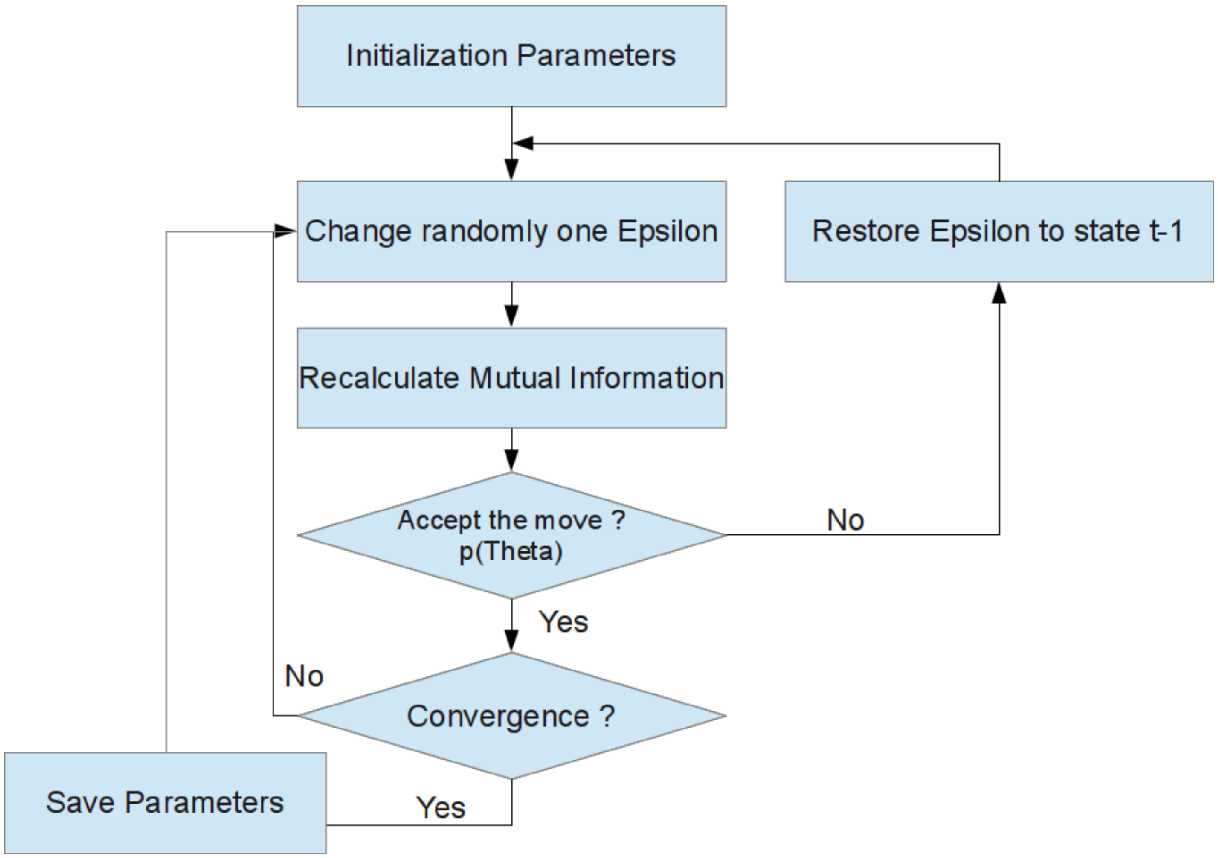

Figure 8: *Monte Carlo Algorithm schematic. The algorithm is initialized with parameters, then it samples the space and accepts moves with a probability defined by the equation 13. Finally if the information saturates then it starts sampling the parameters every 1000 steps.*

4.6 Intra and Inter run variance calculation

The InterRun variance was calculated for each set of parameter:

$$Variance_{Interrun}(\varepsilon_i^p) = \frac{\sum_r^{Runs}(\varepsilon_i^p)_r^2}{Runs} - \frac{(\sum_r^{Runs}(\varepsilon_i^p)_r)^2}{Runs^2} \tag{14}$$

The IntraRun variance was calculated in each run over the time:

$$Variance_{IntraRun}(\varepsilon_i^p) = \frac{\sum_t^T(\varepsilon_i^p)_t^2}{T} - \frac{(\sum_t^T(\varepsilon_i^p)_t)^2}{T^2} \tag{15}$$

4.7 Normalization of Epsilons

Due to the non parametric property of MI, parameters need to be normalized in order to perform any statistics on them. Indeed, MI does not change if the parameters are multiplied by a constant.

In order to normalize them, we first removed the mean values of the epsilons at the position p, and then divided by sum of all epsilons.

$$(\varepsilon_i^p)_{norm} = \frac{\varepsilon_i^p - \frac{\sum_i \varepsilon_i^p}{20}}{\sqrt{\sum_i \sum_p (\varepsilon_i^p)^2}} \tag{16}$$

Moreover, due to the log ratio initial condition, parameters were inferred centered around 0 with positives values being the best energy. In order to be able to fit Boltzmann laws, we reversed the distribution and shifted it to positive value only.

$$(\varepsilon_i^p)_{boltzmann} = -[(\varepsilon_i^p)_{norm} - max(\{\varepsilon_i\}_{norm}^p)] \tag{17}$$

4.8 Boltzmann Probability

The Boltzmann Distribution (also called the Gibbs distribution) describes the probability of a given particle to be in a state i when all micro-states are known. The general rule is written : $p_i = \frac{e^{-\beta E_i}}{Z}$ where $\beta = \frac{1}{k_b T}$ that is one over the Boltzmann Constant (k_b) times the temperature (T) in Kelvin, E_i the energy of the micro-state and $Z = \sum_i e^{-\beta E_i}$ the partition function that is the sum over all the micro-states. This formula is intuitive as the probability of a micro-state is given by it's energy normalized by the sum of energies in the system. Nevertheless it is often easier to measure the probability of a state than it's energy, but it is possible to reverse it in order to find the energy of a state when all states are known.

In our case, we can assume that we have only 2 state: Bound or Unbound. And, that our system is ergodic, that is the average of the time a protein is bound is the same as the average number of bound protein in a population (if the population is big). Thus, the probability of having a protein bound can be written as:

$$p_{bound}(\sigma) = \frac{e^{-\beta E_{bound}}}{e^{-\beta E_{bound}} + e^{-\beta E_{unbound}}} \tag{18}$$

Which can be rewritten as:

$$p_{bound}(\sigma) = \frac{1}{1 + e^{-\beta \Delta E}} \tag{19}$$

where $\Delta E = E_{unbound} - E_{bound}$

In the experiments of Fowler *et al.* [4], they selected the proteins as a function of their states.

If a protein was bound to the ligand, it will pass to the next round, otherwise it is lost. Using the Boltzmann distribution we can then write that the probability of finding the protein at round 1 is :

$$p^{bound}_{round_1}(\sigma) = \frac{1}{1+e^{-\beta\Delta E}} \cdot p^{bound}_{round_0}(\sigma) \tag{20}$$

As the probability only depend on the energy at round k the probability is:

$$p^{bound}_{round_k}(\sigma) = \left(\frac{1}{1+e^{-\beta\Delta E}}\right)^k \cdot p^{bound}_{round_0}(\sigma) \tag{21}$$

Which can be rewritten as the log-ratio of probability:

$$\log\left(\frac{p^{bound}_{round_k}(\sigma)}{p^{bound}_{round_0}(\sigma)}\right) = k \cdot \log(1+e^{-\beta\Delta E}) \tag{22}$$

In the low energy binding limit, when $e^{-\beta\Delta E} >> 1$ then the log-ratio follows a linear law of the form :

$$\log\left(\frac{p^{bound}_{round_k}(\sigma)}{p^{bound}_{round_0}(\sigma)}\right) = -k\beta\Delta E \tag{23}$$

4.9 Protein Visualization

Protein structures where visualized and analyzed using the Jmol software (www.jmol.org). The PDB accession number for the structures are : hYap65: 1JMQ, ABPP1: 2HO2, NEDD4: 2KPZ.

References

[1] A Matouschek, J T Kellis, L Serrano, and A R Fersht. Mapping the transition state and pathway of protein folding by protein engineering. *Nature*, 340(6229):122–6, July 1989.

[2] B C Cunningham and J A Wells. High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. *Science (New York, N.Y.)*, 244(4908):1081–5, June 1989.

[3] Sachdev S Sidhu and Shohei Koide. Phage display for engineering and analyzing protein interaction interfaces. *Current opinion in structural biology*, 17(4):481–7, August 2007.

[4] Douglas M Fowler, Carlos L Araya, Sarel J Fleishman, Elizabeth H Kellogg, Jason J Stephany, David Baker, and Stanley Fields. High-resolution mapping of protein sequence-function relationships. *Nature methods*, 7(9):741–6, September 2010.

[5] Carlos L Araya, Douglas M Fowler, Wentao Chen, Ike Muniez, Jeffery W Kelly, and Stanley Fields. A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proceedings of the National Academy of Sciences of the United States of America*, 109(42):16858–63, October 2012.

[6] Justin B Kinney, Anand Murugan, Curtis G Callan, and Edward C Cox. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proceedings of the National Academy of Sciences of the United States of America*, 107(20):9158–63, May 2010.

[7] Justin B Kinney, Gasper Tkacik, and Curtis G Callan. Precise physical models of protein-DNA interaction from high-throughput data. *Proceedings of the National Academy of Sciences of the United States of America*, 104(2):501–6, January 2007.

[8] E K Koepf, H M Petrassi, G Ratnaswamy, M E Huff, M Sudol, and J W Kelly. Characterization of the structure and function of W –> F WW domain variants: identification of a natively unfolded protein that folds upon ligand binding. *Biochemistry*, 38(43):14338–51, October 1999.

[9] E Krissinel and K Henrick. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica Section D*, 60(12 Part 1):2256–2268, 2004.

[10] Junchao Guo, Jörg Kleeff, Yupei Zhao, Junsheng Li, Thomas Giese, Irene Esposito, Markus W Büchler, Murray Korc, and Helmut Friess. Yes-associated protein (YAP65) in relation to Smad7 expression in human pancreatic ductal adenocarcinoma. *International journal of molecular medicine*, 17(5):761–7, May 2006.

[11] Michele I Vitolo, Ian E Anglin, William M Mahoney, Keli J Renoud, Ronald B Gartenhaus, Kurtis E Bachman, and Antonino Passaniti. The RUNX2 transcription factor cooperates with

the YES-associated protein, YAP65, to promote cell transformation. *Cancer biology & therapy*, 6(6):856–63, June 2007.

[12] W.K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, pages 97–109, 1970.

[13] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–.

[14] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3):90–95, 200.

[15] S. Behnel, R. Bradshaw, C. Citro, L. Dalcin, D.S. Seljebotn, and K. Smith. Cython: The best of both worlds. *Computing in Science Engineering*, 13(2):31 –39, 2011.

5 Acknowledgment

I would like to thanks my supervisors who were more than patient and helped me comprehend notions of physics that would have been really hard to grasp without their help. I would also like to thanks particularly Marc Santolini for its precious day to day help and discussions during the course of that internship. This internship was performed at the ENS ULM, in collaboration between the Laboratoire de Physique Theorique (LPT) and the Laboratoire de Physique Statistique (LPS)