Technische Universität Darmstadt
Department of Physics

Master Thesis

Contact Prediction in Amino Acid Networks With Generalized Entropies

May 25, 2016

Author:
Frederic Hummel

Appraisals:
Prof. Dr. Kay Hamacher
Prof. Dr. Barbara Drossel

Computational Biology And Simulation

Contact Prediction in Amino Acid Networks With Generalized Entropies
Kontaktvorhersage in Aminosäurenetzwerken basierend auf Entropie Verallgemeinerung

Vorgelegte Master-Thesis von Frederic Hummel, Matrikelnummer: 1756528

1. Gutachten: Prof. Dr. Kay Hamacher
2. Gutachten: Prof. Dr. Barbara Drossel

Weitere Betreuung: Michael Schmidt, M.Sc.

Tag der Einreichung:

Technische Universität Darmstadt
Fachbereich Physik
Computational Biology and Simulation
Prof. Dr. Kay Hamacher

Abstract

Computational methods for protein contact prediction have recently attracted great attention due to both laborious and monetary expensiveness of experimental visualization methods of 3D structure. To infer direct interactions from *multiple sequence alignments* (MSA) of proteins, entropy maximizing methods are state-of-the-art. To improve the prediction fidelity and outperform existing algorithms, we enhance the *direct coupling analysis* (DCA) introduced by Weigt et al. We propose a new parameter to generalize the classical Shannon entropy to incorporate the non-equilibrium character biological organisms inhibit not met by the standard assumption of statistical mechanics. With this ansatz, we are able to model the evolutionary selection pressure biological organisms are subjected to. We show that the *mean field inversion* first introduced in DCA is not adoptable using *Tsallis or Rényi entropy*. Therefore, we employ the method of *pseudo likelihood maximization* to solve the *inverse Ising problem*. As our main result we find biases in small sample sizes, such that the Boltzmann-Gibbs distribution is not the appropriate model to match the data anymore, to be well described by Tsallis' distribution. Hence, it is an adequate, analytically motivated substitution for heuristic approaches, such as pseudo counts and reweighting. The proposed algorithm, which is inspired by Ekeberg et al., has computational cost $\mathcal{O}\left(QBL^3\right)$ in the dimension of the MSA $B \times L$ and length of the amino acid alphabet Q.

Zusammenfassung

Computergestützte Methoden zur Vorhersage von Kontakten in Proteinen haben in den letzten Jahren viel Aufmerksamkeit erlangt, weil die ansonsten experimentellen Methoden zur Sichtbarmachung der 3D-Struktur sowohl aufwändig als auch teuer sind. Zur Kontaktvorhersage aufgrund von direkten Zweikörper-Wechselwirkungenen etabliert sich die Entropiemaximierung als die bevorzugte Herangehensweise in der aktuellen Literatur. Um die Güte der Vorhersagen zu erhöhen und die zur Zeit benutzten Algorithmen in ihrer Leistung zu übertreffen, verallgemeinern wir die von Weigt et al. vorgeschlagene *Direct Coupling Analysis* (DCA), indem wir die klassische Shannon-Entropie um einen Parameter erweitern. Dieser bezieht den Nicht-Gleichgewichtscharakter, den ein biologischer Organismus aufweist, der aber durch die übliche Gleichgewichtsthermodynamik nicht berücksichtigt wird, mit ein. Durch diesem Ansatz haben wir die Möglichkeit den evolutionären Selektionsdruck, dem ein biologischer Organismus ausgesetzt ist, zu modellieren. Wir zeigen, dass die Korrelationsmatrix-Inversion des Mean-Field-Ansatzes nicht unter unserer *Entropieverallgemeinerung nach Tsallis oder Rényi* umsetzbar ist. Daher benutzen wir das *Pseudo-Likelihood-Schätzverfahren* um das *inverse Ising-Problem* zu lösen. Als zentrales Ergebnis dieser Arbeit zeigt sich, dass sich die Verzerrung durch kleine Datensätze, der Art, dass die Boltzmann-Gibbs Verteilung zur Modellierung nicht mehr geeignet ist, sehr gut durch die Tsallis Verteilung beschreiben lässt. Daher ist sie ein geeigneter, analytisch motivierter Ersatz um heuristische Ansätze, wie Pseudo Counts und Reweighting abzulösen. Der von uns vorgeschlagene, von Ekeberg et al. inspirierte Algorithmus hat den Rechnungsaufwand $\mathcal{O}\left(QBL^3\right)$ bezogen auf die Dimension des herangezogenen *Multiple Sequence Alignments* $B \times L$ und die Anzahl der Aminosäuren Q.

Declaration of Originality

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. All citations from literature or the Internet are marked as such. I have documented all methods, data, and processes truthfully and I have not manipulated any data.

Date	Frederic Hummel

Contents

1 Introduction

From the point of a physicist a protein may be nothing more than a folded one-dimensional chain of amino acids, or, to get to the heart of it, a one-dimensional lattice of states, taken from an alphabet consisting of the 20 amino acids, and in some cases also including a gap. Since at protein production in a biological organism, the chain is created element-wise and folds upon full or partial completion, we have reason to argue that only the positional arrangement of states is accountable for the proteins 3D structure, which is essential to its function. From the residual setup we try to infer the protein's final folding by identifying those amino acids which are evolutionary correlated and thus, in direct spatial contact.
We use biological data in form of *multiple sequence alignments* (MSA), which are matrices containing the states of an amino acid chain element-wise, each column representing a residue along the protein and each row representing a protein taken from different organisms. We are able to identify relevant residues of amino acids accountable for the protein's function by their residual conservation throughout the alignment. Such data is vastly available for example at [Fin+13; Tho+97; Jon+12]. The idea behind this is that if an amino acid in the chain is substituted, due to mutation, destabilizing the proteins structure, the amino acid in contact with the former has to adjust in order to maintain the protein's function. This implies a strong residual coupling between structure-relevant amino acids. Thus, our aim is to identify high two-point correlations in MSAs. In order to do so and to implicitly exclude other kinds of couplings like indirect correlations, the tool of entropy maximization was introduced in biological data analysis (reviewed in [SMS15]), where the empirical single-site and pair frequency counts of the considered MSA are incorporated as constraints on the entropy function.
Maximizing the entropy yields a *probability distribution* (PD) entailing the familiar Boltzmann-Gibbs factor and therein the Hamiltonian of a generalized Potts-Model in which every residue can in principle interact with every other residue along the chain depending on their states (Section 1.2). Computing the couplings appearing in the governing Hamiltonian from the mesoscopic frequency counts is called the *inverse Ising problem* and it requires calculation of the partition function of the physical system. Due to the length of both the alphabet (21 states) and of typical proteins (ca. 100 amino acids) this is computational expensive (scaling exponentially with the length) and approximations have to be employed.
To improve the predictive power of this approach we want to take into account the fact that proteins are characterized by a high complexity emerging in living (i.e. non-equilibrium) environments, a fact that may not be regarded when using statistical mechanics originally introduced as a short-range interaction gas theory. Therefore our ansatz is to generalize the notion of entropy giving space for deviations from equilibrium in form of evolutionary pressure.
Rényi first introduced a possible generalization of Shannon entropy in a sense of generalizing the notion of expectation value to non-linearity. (Note that entropy may be interpreted as the expected value of the information gained given the occurrence of a micro-state the physical system assumes.) He showed that only one entropy function apart from the usual Shannon entropy retains the property of additivity [Rén66] (Section 2.2). Another generalization was introduced by Tsallis, who also considered the notion of information gain itself [Tsa88] (Section 2.1). Both generalizations introduce a reweighting factor to the probabilities of assuming a specific micro-state in form of an exponent. That way it is possible to tune the importance of likely events relative to unlikely events. We may now narrow or widen the peaks of distributions to model a high or low selection pressure, respectively. In addition, the generalization parameter gives way to avoid heuristic distribution reshaping like pseudo counts or reweighting factors, which values are only determined phenomenologically to receive optimal results. This, however, does not suffice as a requirement on a serious protein model.

Recently the method of pseudo likelihood maximization has attracted a lot of attention as a technique to optimally fit data to suggested theoretical models [DRT14]. We also employ this approximate approach (in Section 3.2), because the newly attained probability distributions lack a stringent linearized mean field description (as shown in Section 3.1).
With this theoretical background an algorithm to solve the inverse Ising problem will be presented (Section 4.1) and thoroughly tested on a computationally tractable toy model (Section 4.2). Let us introduce incrementally the theoretic framework this work is based on.

1.1 Proteins

In our analysis we do not consider proteins in their 3D structure but as biological data given in the form of *multiple sequence alignments* (MSA) (compare Figure 1). An MSA is a collection of letters in matrix form, each letter representing one of 20 amino acids or a gap, and each row constitutes a protein such that each column gives positional information and we may compare residual conservation among proteins. Sequence statistics are attained by column-wise counting of occurring states, which we call spins from now on referring to the Ising model, and by pair-wise counting, respectively

$$f_i(k) = \frac{1}{B_{\text{eff}}} \sum_{b=1}^{B} \frac{1}{m_b} \delta_{\sigma_i^{(b)} k}, \tag{1.1}$$

$$f_{ij}(k,l) = \frac{1}{B_{\text{eff}}} \sum_{b=1}^{B} \frac{1}{m_b} \delta_{\sigma_i^{(b)} k} \delta_{\sigma_j^{(b)} l}, \tag{1.2}$$

where $f_i(k)$ denotes the single-site frequency count of spin k at position i and $f_{ij}(k,l)$ denotes the pair frequency count of spin k at position i and spin l at position j. For indices $i,j \in \{1,\dots,L\}$, with L being the length of the amino acid chain, i.e. the number of columns in the MSA, and $k,l \in \{1,\dots,Q\}$, with Q being the number of spins a residue is able to assume. Throughout this work Q will in general be 21 if not denoted otherwise. b counts the rows of the MSA overall containing B proteins, δ denotes the Kronecker symbol and $\sigma_i^{(b)}$ is the spin at residue i in the b-th protein of the MSA.

$$m_b = \left| \{ a \in \{1,\dots,B\} : \text{similarity}(\boldsymbol{\sigma}^{(a)}, \boldsymbol{\sigma}^{(b)}) \geq x \} \right| \tag{1.3}$$

is the number of sequences $\boldsymbol{\sigma}$ in the MSA that are considered similar, since more than a fraction $0 \leq x \leq 1$ of residues carry the same spin. Thus $1/m_b$ becomes a reweighting factor in order to correct biases from sequencing. This is a heuristic approach and throughout the literature $x = 0.8$ has shown to provide good results. $B_{\text{eff}} = \sum_b 1/m_b$ is then the effective number of sequences in the MSA. In the literature further phenomenological reshapinglike pseudo counts have been introduced which then turned out to be crucial for prediction fidelity [Mor+11]. It is our aim to show those ansatzes to be superfluous. We then set $m_b = 1$, thus $B_{\text{eff}} = B$, and use the Tsallis parameter q, which will be introduced in 2.1, instead to quantify possible biases in the distribution.

1.2 Direct Coupling Analysis

In order to attain a global statistical model describing an MSA that only inhibits one-point and two-point correlations we will combine the Lagrangian method with *maximum entropy inference* [Jay57a; Jay57b] including the empirical findings from Section 1.1 as constraints on the PD using Lagrange multipliers. This has recently become a common tool in bioinformatics [MM09; Wei+09; CM11; Mor+11; AE12; NB12a].

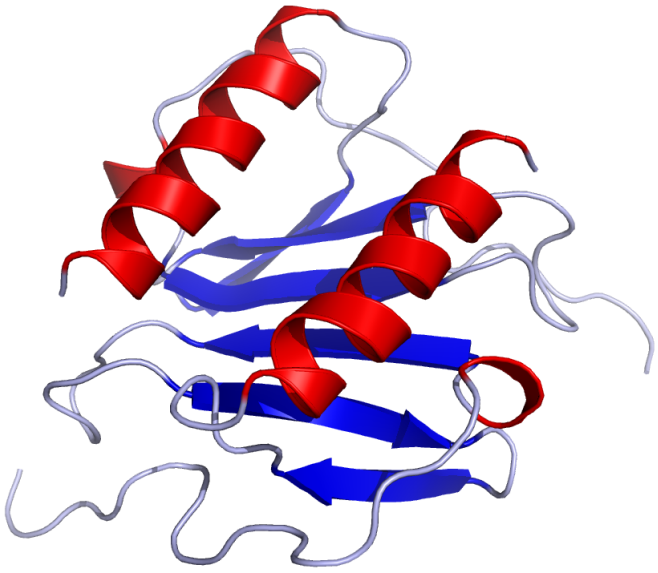

Figure 1: Schematic image of two Interleukin8 (CXCL8) proteins [Com06], a signaling complex. It is a folded one-dimensional chain (grey) with secondary structure highlighted in red (α-helices) and blue (β-strands). Amino acids are in physical contact not only with their chain neighbors but also due to the fold, for example vertically at the horizontal β-strands.

The Shannon entropy [Sha48] reads

$$S = \sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma}) \ln \frac{1}{p(\boldsymbol{\sigma})}, \tag{1.4}$$

where possible constants (like Boltzmann's k and such arising from using another logarithm base than the natural one) are set to unity throughout this study. Furthermore $p(\boldsymbol{\sigma})$ is the probability of the occurrence of sequence $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_i, \ldots, \sigma_L)$ and σ_i denotes the spin at residue i. The constraints read

$$\sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma}) \stackrel{!}{=} 1, \tag{1.5}$$

$$\sum_{\boldsymbol{\sigma}|\sigma_i=k} p(\boldsymbol{\sigma}) \equiv p_i(k) \stackrel{!}{=} f_i(k), \tag{1.6}$$

$$\sum_{\boldsymbol{\sigma}|\substack{\sigma_i=k \\ \sigma_j=l}} p(\boldsymbol{\sigma}) \equiv p_{ij}(k,l) \stackrel{!}{=} f_{ij}(k,l), \tag{1.7}$$

where the first one is necessary for normalization. For an MSA with length L and Q possible states that makes $Q \cdot L$ equations of type (1.6), each incorporated with Lagrange multipliers $h_i(k)$; and $Q^2 \cdot L(L-1)/2$ equations of type (1.7), each incorporated with Lagrange multipliers $J_{ij}(k,l)$. The latter is due to the symmetry of $f_{ij}(k,l)$ and $f_{ji}(l,k)$ and we omit redundancy such that only $i < j$. Maximizing the entropy under these constraints leads to the distribution

$$p(\boldsymbol{\sigma}) = \frac{1}{Z} \exp\{-H(\boldsymbol{\sigma})\}, \tag{1.8}$$

with the Hamiltonian

$$H(\boldsymbol{\sigma}) = -\sum_i h_i(\sigma_i) - \sum_{i<j} J_{ij}(\sigma_i, \sigma_j), \tag{1.9}$$

where $h_i(\sigma_i)$ and $J_{ij}(\sigma_i, \sigma_j)$ are the Lagrange multipliers for spin σ_i at position i and σ_j at position j, according to the alignment $\boldsymbol{\sigma}$. The partition function reads

$$Z = \sum_{\boldsymbol{\sigma}} \exp\{-H(\boldsymbol{\sigma})\}. \tag{1.10}$$

This is a Boltzmann distribution for unit temperature and the well-known Hamiltonian of the Potts model [Pot52]. In terms of statistical physics we can thus interpret the Lagrange multipliers $\boldsymbol{h_i}$ as local fields governing the single residues at position i within the network and $\boldsymbol{J_{ij}}$ as two-point interactions between residue i and j. In the following we also write $\hat{\boldsymbol{h}}$ and $\hat{\boldsymbol{J}}$ including the fields and couplings of all residues in one expression.
The distribution can in principle be calculated, but it requires calculation of the partition function, which has exponential computational cost of order $\mathcal{O}(Q^L)$.

1.3 Mean Field Approximation

When assuming the two-point interactions to be small, calculation of the entire partition function can be avoided [Ple82; GY91; SM09]. For biological sequence data this method has been introduced by [Mor+11] and further investigated by [NB12a; NB12b; RT12]. This method is called mean-field approximation due to the fact that with this assumption the system is already relatively well described only by the local fields. A review of common approximation techniques originally introduced for neural networks and their performance can be found in [RAH09]. Introducing the perturbed Hamiltonian

$$H_\alpha(\boldsymbol{\sigma}) = -\sum_i h_i(\sigma_i) + \alpha H_I(\boldsymbol{\sigma}), \tag{1.11}$$

with perturbation parameter $0 \leq \alpha \leq 1$ and the interaction Hamiltonian

$$H_I(\boldsymbol{\sigma}) = -\sum_{i<j} J_{ij}(\sigma_i, \sigma_j), \tag{1.12}$$

we find the Gibbs potential associated with the Helmholtz free energy $F = -\ln Z$ upon Legendre transformation with respect to the fields $h_i(k)$

$$G(\alpha) = -\ln Z(\alpha) + \sum_{i,k} h_i(k) p_i(k). \tag{1.13}$$

Due to the properties of the Legendre transformation we know [ZRM09]

$$h_i(k) = \frac{\partial G}{\partial p_i(k)}, \tag{1.14}$$

$$(C^{-1})_{ij}(k,l) = \frac{\partial^2 G}{\partial p_i(k) \partial p_j(l)}, \tag{1.15}$$

where C is the correlation matrix

$$C_{ij}(k,l) = p_{ij}(k,l) - p_i(k) p_j(l) \tag{1.16}$$

$$= f_{ij}(k,l) - f_i(k) f_j(l), \tag{1.17}$$

which due to the second equality is directly accessible from the MSA. When Taylor expanding the Gibbs potential around $\alpha = 0$ up to linear order we find

$$G(0) = \sum_{i,k} p_i(k) \ln p_i(k), \tag{1.18}$$

which follows from the Legendre transform for $\alpha = 0$ removing the average energy, which leaves us with the negative entropy of uncoupled spins. Furthermore

$$\frac{\mathrm{d}G(\alpha)}{\mathrm{d}\alpha} = -\frac{\mathrm{d}}{\mathrm{d}\alpha} \ln Z(\alpha) \tag{1.19}$$

$$= -\sum_{\boldsymbol{\sigma}} H_I(\boldsymbol{\sigma}) \frac{\exp\{-H_\alpha(\boldsymbol{\sigma})\}}{Z(\alpha)} \tag{1.20}$$

$$= -\langle H_I \rangle_\alpha \,. \tag{1.21}$$

The last term denotes the expected value of the couplings for a given sequence with respect to α. At $\alpha = 0$ the joint distribution of all variables factorizes to single-site distributions:

$$\left.\frac{\mathrm{d}G(\alpha)}{\mathrm{d}\alpha}\right|_{\alpha=0} = -\sum_{i<j} \sum_{k,l} J_{ij}(k,l)\ p_i(k) p_j(l). \tag{1.22}$$

Using Equation (1.15) we obtain

$$(C^{-1})_{ij}(k,l) = -J_{ij}(k,l) \tag{1.23}$$

for $i \neq j$. Thus, the problem simplifies to inversion of the correlation matrix, which can be taken directly from the empirical data. Note that C is not invertible as it is a $Q \cdot L \times Q \cdot L$ matrix with rank $Q \cdot (L-1) \times Q \cdot (L-1)$. This is due to the fact that for the marginal distributions we have

$$\sum_l f_{ij}(k,l) = f_i(k), \tag{1.24}$$

$$\sum_k f_{ij}(k,l) = f_j(l), \tag{1.25}$$

$$\sum_{k,l} f_{ij}(k,l) = \sum_k f_i(k) = \sum_l f_j(l) = 1, \tag{1.26}$$

reducing the degrees of freedom. An appropriate gauge can be found to reduce the dimension of the problem or methods for matrix pseudo inversion have to be employed. Both can be performed easily. For randomly distributed spins within the MSA the method is exact [Bar+14].

1.4 Likelihood as Learning Criterion

An ansatz for exact results of *direct coupling analysis* (DCA) is the maximum likelihood principle for model learning, which for numerical reasons practically means to minimize the negative logarithmic likelihood function l over a set of parameters $\boldsymbol{\theta} = \{\hat{\boldsymbol{h}}, \hat{\boldsymbol{J}}\}$ for each observation $\boldsymbol{\sigma}$:

$$l(\boldsymbol{\theta}, \boldsymbol{\sigma}) = -\log p(\boldsymbol{\sigma}; \boldsymbol{\theta}), \tag{1.27}$$

where $p(\boldsymbol{\sigma}; \boldsymbol{\theta})$ is the probability of occurrence of sequence $\boldsymbol{\sigma}$, given $\boldsymbol{\theta}$. For an MSA with B observations and when $p(\boldsymbol{\sigma}; \boldsymbol{\theta})$ is given by Equation (1.8) we find

$$l(\hat{\boldsymbol{h}}, \hat{\boldsymbol{J}}) = -\frac{1}{B_{\mathrm{eff}}} \sum_{b=1}^{B} \frac{1}{m_b} \ln \left\{ \frac{1}{Z} \exp\left\{ -H(\boldsymbol{\sigma}^{(b)}) \right\} \right\}$$

$$= \ln Z - \sum_i \sum_k h_i(k) f_i(k) - \sum_{i<j} \sum_{k,l} J_{ij}(k,l) f_{ij}(k,l). \tag{1.28}$$

This approach still requires knowledge of the partition function, which is not feasible due to the sheer size of the biological systems under investigation. Therefore we substitute the probability in Equation (1.27) with the conditional probability of finding spin σ_r at residue r given the spins in all other residues as introduced for DCA in [Eke+13; EHA14] as

$$\begin{aligned} p(\sigma_r|\boldsymbol{\sigma}_{\setminus r}) &= \frac{p(\sigma_r \cup \boldsymbol{\sigma}_{\setminus r})}{p(\boldsymbol{\sigma}_{\setminus r})} \\ &= \frac{e^{h_r(\sigma_r)+\sum_{i\neq r} J_{ri}(\sigma_r,\sigma_i)}}{\sum_k e^{h_r(k)+\sum_{i\neq r} J_{ri}(k,\sigma_i)}}. \end{aligned} \tag{1.29}$$

This will lead to less accurate predictions than likelihood maximization, but is not dependent on the partition function and therefore computable without further approximations. With this substitution, we introduce the (negative and logarithmic) pseudo likelihood

$$l_{\text{pseudo}}(\hat{\boldsymbol{h}}, \hat{\boldsymbol{J}}) = \sum_{r=1}^{L} \left\{ z_r - \sum_k h_r(k) f_r(k) - \sum_{i>r} \sum_{k,l} J_{ri}(k,l) f_{ri}(k,l) \right\}, \tag{1.30}$$

where we use for $i < r$ the symmetry of $J_{ir}(l,k) = J_{ri}(k,l)$, where

$$z_r = \frac{1}{B_{\text{eff}}} \sum_{b=1}^{B} \frac{1}{m_b} \ln \sum_k e^{h_r(k)+\sum_{i\neq r} J_{ri}(k,\sigma_i^{(b)})} \tag{1.31}$$

is a position specific constant. Due to the large number of parameters of this model, to avoid over-fitting, we need to introduce a regularization scheme

$$R(\hat{\boldsymbol{h}}, \hat{\boldsymbol{J}}) = \lambda_h \sum_i \|\boldsymbol{h_i}\|_2^2 + \lambda_J \sum_{i<j} \left\|\boldsymbol{J_{ij}}\right\|_2^2, \tag{1.32}$$

where λ_h and λ_J are parameters. This is an ℓ_2 regularization. In principle, other regularizers like ℓ_1 investigated in [RWL+10] are also possible, but the convex, differentiable ℓ_2 is the most convenient choice. Thus, the corresponding fields and couplings to a given MSA will be obtained by posing

$$\{\hat{\boldsymbol{h}}^\star, \hat{\boldsymbol{J}}^\star\} = \arg\min_{\hat{\boldsymbol{h}},\hat{\boldsymbol{J}}} \left\{ l_{\text{pseudo}}(\hat{\boldsymbol{h}}, \hat{\boldsymbol{J}}) + R(\hat{\boldsymbol{h}}, \hat{\boldsymbol{J}}) \right\}. \tag{1.33}$$

The computational cost of an algorithm to solve this optimization problem is $\mathcal{O}(BL^2Q)$, for the alignment size $B \times L$ and a Q-letter alphabet.

2 Entropy

Shannon introduced entropy as the expected value of the information gained when finding a physical system in a certain micro-state $\boldsymbol{\sigma}$

$$S \equiv \langle I \rangle = \sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma}) I(\boldsymbol{\sigma}), \quad I(\boldsymbol{\sigma}) = -\ln p(\boldsymbol{\sigma}). \tag{2.1}$$

The information gain $I(\boldsymbol{\sigma})$, when finding the system in micro-state $\boldsymbol{\sigma}$, is the negative logarithm of the probability of occurrence $p(\boldsymbol{\sigma})$, which is the same as the logarithm of the surprise $1/p(\boldsymbol{\sigma})$ corresponding to $\boldsymbol{\sigma}$. In the following sections we will generalize both, the notion of information gain and the notion of expected value. The latter is usually given in the linear form for an arbitrary operator $O(\boldsymbol{\sigma})$ with appropriate domain as

$$\langle O \rangle = \sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma}) O(\boldsymbol{\sigma}). \tag{2.2}$$

2.1 Tsallis' Generalization

In 1988 Constantino Tsallis proposed an new entropy function [Tsa88]

$$S_{\mathrm{T}} = \frac{1}{1-q} \left(\sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma})^q - 1 \right). \tag{2.3}$$

It recovers Shannon's entropy, as we shall see further below, in the limit

$$\lim_{q \to 1} S_{\mathrm{T}} = S. \tag{2.4}$$

Inspection of the equation reveals that it can be interpreted as generalizations of both, the information gain and the expected value [Mas05]. To show this we define the q-information gain as

$$I_q(\boldsymbol{\sigma}) = \frac{1 - p(\boldsymbol{\sigma})^{1-q}}{1-q}, \tag{2.5}$$

and the q-expectation value for an appropriate operator $O(\boldsymbol{\sigma})$ as

$$\langle O \rangle_q = \sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma})^q O(\boldsymbol{\sigma}). \tag{2.6}$$

The relation to the original entropy function becomes evident through some simple algebra: Consider the limit-definition of the exponential function

$$e^x = \lim_{n \to \infty} \left(1 + \frac{x}{n}\right)^n. \tag{2.7}$$

The Tsallis parameter substitutes n with a variable with the same quality but in the limit approaching 1 instead of ∞ in the way that $q = 1 - \frac{1}{n}$, defining the q-generalized exponential function by omitting the taking of the limit

$$e_q^x = (1 + (1-q)x)^{\frac{1}{1-q}}. \tag{2.8}$$

This induces its inverse function, the q-generalized logarithm

$$\ln_q y = \frac{y^{1-q} - 1}{1-q}. \tag{2.9}$$

Hence the Tsallis information gain can also be written as the negative q-generalized logarithm of the probability of occurrence

$$I_q(\boldsymbol{\sigma}) = -\ln_q p(\boldsymbol{\sigma}), \tag{2.10}$$

and the Tsallis entropy is found to be its q-expectation value

$$S_\mathrm{T} = -\sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma})^q \ln_q p(\boldsymbol{\sigma}). \tag{2.11}$$

under the assumption that $\sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma}) = 1$. In the framework of the Lieb-Ingvason approach of statistical mechanics [LY99] it is the axiom of additivity of entropy that we have to give up to allow for the Tsallis generalization. Due to the logarithmic form of the original term of information gain it is additive regarding two independent events. In the Tsallis case the additivity relation is replaced by

$$S_\mathrm{T}(A+B) = S_\mathrm{T}(A) + S_\mathrm{T}(B) + (1-q)S_\mathrm{T}(A)S_\mathrm{T}(B), \tag{2.12}$$

for two independent systems A and B. This is a desired property since highly correlated systems do not necessarily inhibit extensive entropy in the form we know from equilibrium thermodynamics.

2.2 Rényi's Generalization

The second entropy generalization we consider in this work is introduced by Alfred Rényi in 1966 [Rén66]

$$S_\mathrm{R} = \frac{1}{1-q} \ln \sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma})^q. \tag{2.13}$$

Rényi's aim was to discover under which generalization the additivity of the entropy is preserved and found a possibility in the definition of the expected value, which he first considered as

$$\langle O \rangle_f = f^{-1}\left(\sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma}) f(O(\boldsymbol{\sigma})) \right), \tag{2.14}$$

where f is an arbitrary one-dimensional function. Rényi proved that the entropy additivity for an additive information gain as $I(\boldsymbol{\sigma}) = -\ln p(\boldsymbol{\sigma})$ is preserved only for one choice of f, beside the linear function, being [Mas05]

$$f(x) = a\ e^{(1-q)x} + b, \tag{2.15}$$

with constants a and b, which reduce in the calculation of actual expected values. For the limit of $q \to 1$ the exponential function can be expanded and a linear function of x is recovered, thus retrieving the standard linear expected value. With the above function (2.15) we find for appropriate operators $O(\boldsymbol{\sigma})$

$$\langle O \rangle_e = \frac{1}{1-q} \ln \left(\sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma}) e^{(1-q)O(\boldsymbol{\sigma})} \right). \tag{2.16}$$

We call the Rényi entropy the exponential expected value of the classical information gain, whereas the Tsallis entropy is the q-expected value of the q-generalized information gain:

$$S_\mathrm{T} = \langle I_q(\boldsymbol{\sigma}) \rangle_q, \tag{2.17}$$

$$S_\mathrm{R} = \langle I(\boldsymbol{\sigma}) \rangle_e. \tag{2.18}$$

Note that both entropies contain the probabilities of a certain event weighted with the exponent q. In the micro-canonical ensemble the equal probability distribution is preserved but apart from that ensemble the PD will be altered in various ways. This is also dependent on the way the constraints are incorporated as will be shown in the next section.

2.3 Constraints

After introducing his entropy, Tsallis discusses various approaches to include constraints for entropy extremization methods, which are all based on different expectation values of physical operators [TMP98]. To be physically consistent, we think that our above introduced definition of entropy automatically gives rise to only one choice of constraints, namely the one being consistent with the expected value used in the definition of entropy[1].

2.3.1 Linear Constraints

In the case of Rényi entropy we need to incorporate the constraints as exponential expectation values of Kronecker-δ operators. Due to the high complexity of the associated equation we Taylor expand these equations in $\sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma})\delta_{\sigma_i k} \leq 1$ to linear order, so that for q being in the same magnitude as 1 we find (compare Appendix B.1)

$$\left\langle \delta_{\sigma_i k} \right\rangle_e \approx \sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma})\delta_{\sigma_i k} \tag{2.19}$$

to be still a good approximation. We thus derive linear constraints, namely

$$1 = \langle \mathbb{1} \rangle = \sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma})\mathbb{1}, \tag{2.20}$$

$$f_i(k) = \left\langle \delta_{\sigma_i k} \right\rangle = \sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma})\delta_{\sigma_i k}, \tag{2.21}$$

$$f_{ij}(k,l) = \left\langle \delta_{\sigma_i k}\delta_{\sigma_j l} \right\rangle = \sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma})\delta_{\sigma_i k}\delta_{\sigma_j l}. \tag{2.22}$$

We have all the constituents of the Lagrange function we want to extremize in order to acquire a probability distribution at hand now. For better overview we differentiate every part one by one:
Entropy:

$$\frac{\partial}{\partial p(\boldsymbol{\sigma})} S_{\mathrm{R}} = \frac{q}{1-q} \frac{p(\boldsymbol{\sigma})^{q-1}}{\sum_{\boldsymbol{\sigma}'} p(\boldsymbol{\sigma}')^q} \tag{2.23}$$

Normalization, with the associated Lagrange multiplier κ:

$$\frac{\partial}{\partial p(\boldsymbol{\sigma})} \kappa \left(\sum_{\boldsymbol{\sigma}'} p(\boldsymbol{\sigma}') - 1 \right) = \kappa \tag{2.24}$$

Single-site frequency counts, with the associated Lagrange multipliers $h_i(k)$:

$$\frac{\partial}{\partial p(\boldsymbol{\sigma})} \sum_i \sum_k h_i(k) \left(\sum_{\boldsymbol{\sigma}'} p(\boldsymbol{\sigma}')\delta_{\sigma_i k} - f_i(k) \right) = \sum_i h_i(\sigma_i) \tag{2.25}$$

Pair frequency counts, with the associated Lagrange multipliers $J_{ij}(k,l)$:

$$\frac{\partial}{\partial p(\boldsymbol{\sigma})} \sum_{i<j} \sum_{k,l} J_{ij}(k,l) \left(\sum_{\boldsymbol{\sigma}'} p(\boldsymbol{\sigma}')\delta_{\sigma_i k}\delta_{\sigma_j l} - f_{ij}(k,l) \right) = \sum_{i<j} J_{ij}(\sigma_i, \sigma_j) \tag{2.26}$$

The summations go over every residue i, j and every spin k, l. Due to the symmetry of $f_{ij}(k,l) = f_{ji}(l,k)$ only $i < j$ are regarded. By reminding of the Hamiltonian

$$H(\boldsymbol{\sigma}) = -\sum_i h_i(\sigma_i) - \sum_{i<j} J_{ij}(\sigma_i, \sigma_j), \tag{2.27}$$

[1] For further information and detailed calculations on this matter one may also consult the proposal for this work.

we thus find

$$\frac{q}{1-q}\frac{p(\boldsymbol{\sigma})^{q-1}}{\sum_{\boldsymbol{\sigma}'} p(\boldsymbol{\sigma}')^q} + \frac{\kappa}{1-q} - H(\boldsymbol{\sigma}) \stackrel{!}{=} 0. \tag{2.28}$$

To eliminate the parameter κ from the equation we use a technique introduced in [DS+99]. Multiplication of (2.28) with $p(\boldsymbol{\sigma})$ and subsequent summation over all possible sequences yields

$$\frac{q}{1-q} + \frac{\kappa}{1-q} - \langle H \rangle = 0, \tag{2.29}$$

with the expected energy of the distribution

$$\langle H \rangle = \sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma}) H(\boldsymbol{\sigma}), \tag{2.30}$$

$$= -\sum_i \sum_k h_i(k) p_i(k) - \sum_{i<j} \sum_{k,l} J_{ij}(k,l) p_{ij}(k,l) \tag{2.31}$$

where the equality can be calculated using the constraints (2.20) - (2.22). Upon subtraction of (2.28) and (2.29) we eliminate κ. Solving for $p(\boldsymbol{\sigma})$ leads to the PD

$$p(\boldsymbol{\sigma}) = \frac{1}{Z}\left(1 - \frac{q-1}{q}\left\{H(\boldsymbol{\sigma}) - \langle H \rangle\right\}\right)^{\frac{1}{q-1}}, \tag{2.32}$$

with the partition function

$$Z = \left(\sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma})^q\right)^{\frac{1}{1-q}} \tag{2.33}$$

$$= \sum_{\boldsymbol{\sigma}} \left(1 - \frac{q-1}{q}\left\{H(\boldsymbol{\sigma}) - \langle H \rangle\right\}\right)^{\frac{1}{q-1}}. \tag{2.34}$$

The same linear constraints could be employed to calculate a PD based on Tsallis entropy which would in fact be the same up to a constant scaling factor of the form Z^{q-1}. As has been shown in [FMP05a] the derived distributions are equivalent, with and without the factor because it can be absorbed into the Lagrange multipliers. Before we discuss the Rényi probability distribution function any further we consider another approach for the Tsallis entropy.

2.3.2 q Constraints

Our formulation of Tsallis entropy gave rise to what we call the q-expectation value. This leads to a different employment of constraints and thus a different probability distribution, which will be attained in the following. Note that the naive adaption of the newly acquired q-expectation value leads to contradictions in case of the unitary operator

$$\langle \mathbb{1} \rangle = \sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma})^q \mathbb{1} \neq 1 \quad \text{for} \quad q \neq 1, \tag{2.35}$$

which is why we add a normalization. For an appropriate operator $O(\boldsymbol{\sigma})$ we define

$$\langle O \rangle_q = \frac{\sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma})^q O(\boldsymbol{\sigma})}{\sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma})^q}. \tag{2.36}$$

This ensures the expected value of the unity operator to be one. Transposition of the equation gives rise to the interpretation of substituting the operator $O(\boldsymbol{\sigma})$ by its so called centralized operator [Mar+00]

$$O(\boldsymbol{\sigma}) \to O(\boldsymbol{\sigma}) - \langle O \rangle_q \,. \tag{2.37}$$

We then write the q-expectation value as

$$\sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma})^q \left(O(\boldsymbol{\sigma}) - \langle O \rangle_q \right) = 0. \tag{2.38}$$

With the definitions (2.36) and (2.38) the constraints read

$$\sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma})^q \left(\delta_{\sigma_i k} - f_i(k) \right) = 0, \tag{2.39}$$

$$\sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma})^q \left(\delta_{\sigma_i k} \delta_{\sigma_j l} - f_{ij}(k,l) \right) = 0. \tag{2.40}$$

Incorporating these constraints with Lagrange multipliers in the same manner as before and differentiation with respect to $p(\boldsymbol{\sigma})$ leads to

$$\begin{aligned} &\frac{\partial}{\partial p(\boldsymbol{\sigma})} \sum_i \sum_k h_i(k) \sum_{\boldsymbol{\sigma}'} p(\boldsymbol{\sigma}')^q \left(\delta_{\sigma_i k} - f_i(k) \right) \\ &\qquad = q p(\boldsymbol{\sigma})^{q-1} \left(\sum_i h_i(\sigma_i) - \sum_i \sum_k h_i(k) f_i(k) \right), \end{aligned} \tag{2.41}$$

$$\begin{aligned} &\frac{\partial}{\partial p(\boldsymbol{\sigma})} \sum_{i<j} \sum_{k,l} J_{ij}(k,l) \sum_{\boldsymbol{\sigma}'} p(\boldsymbol{\sigma}')^q \left(\delta_{\sigma_i k} \delta_{\sigma_j l} - f_{ij}(k,l) \right) \\ &\qquad = q p(\boldsymbol{\sigma})^{q-1} \left(\sum_{i<j} J_{ij}(\sigma_i, \sigma_j) - \sum_{i<j} \sum_{k,l} J_{ij}(k,l) f_{ij}(k,l) \right). \end{aligned} \tag{2.42}$$

Furthermore the derivative of the Tsallis entropy reads

$$\frac{\partial}{\partial p(\boldsymbol{\sigma})} S_{\mathrm{T}} = \frac{q}{1-q} p(\boldsymbol{\sigma})^{q-1}. \tag{2.43}$$

With the expected energy of the alignment

$$\langle H \rangle = -\sum_i \sum_k h_i(k) f_i(k) - \sum_{i<j} \sum_{k,l} J_{ij}(k,l) f_{ij}(k,l) \tag{2.44}$$

we deduce the following equation:

$$\left(\frac{1}{1-q} - H(\boldsymbol{\sigma}) + \langle H \rangle \right) q p(\boldsymbol{\sigma})^{q-1} - \kappa \stackrel{!}{=} 0, \tag{2.45}$$

where κ is again the Lagrange multiplier associated with the probability normalization. We obtain

$$p(\boldsymbol{\sigma}) = \frac{1}{Z} \left(1 - (1-q) \left\{ H(\boldsymbol{\sigma}) - \langle H \rangle \right\} \right)^{\frac{1}{1-q}} \tag{2.46}$$

upon elimination of κ as

$$\left(\frac{1-q}{q} \kappa \right)^{\frac{1}{1-q}} = \sum_{\boldsymbol{\sigma}} \left(1 - (1-q) \left\{ H(\boldsymbol{\sigma}) - \langle H \rangle \right\} \right)^{\frac{1}{1-q}} = Z. \tag{2.47}$$

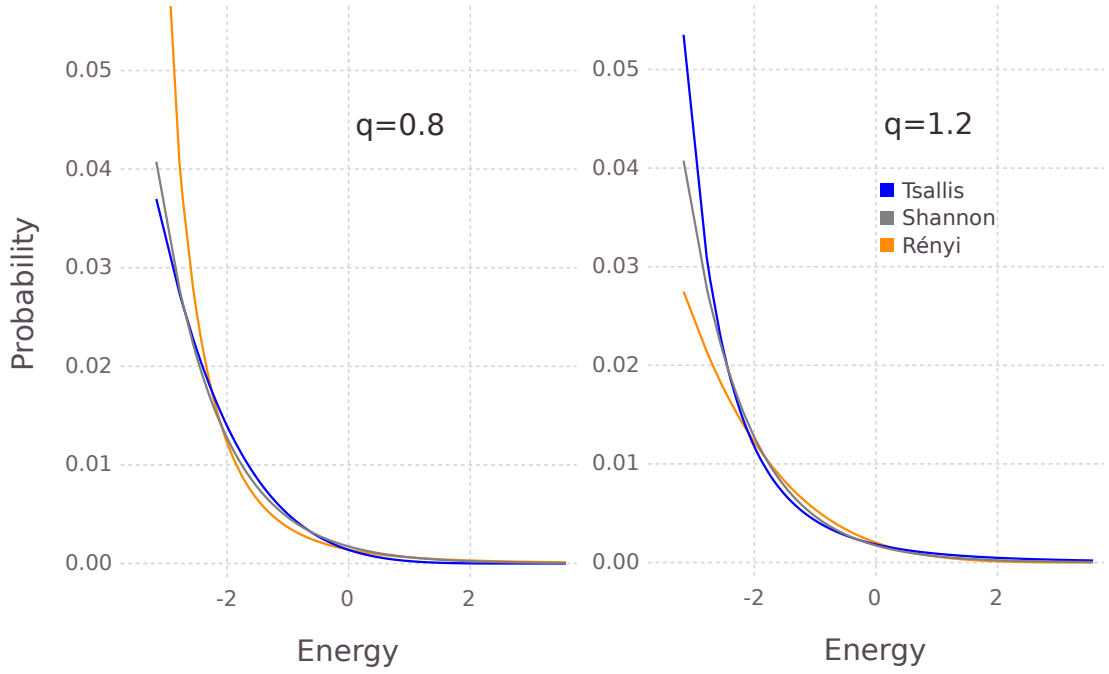

Figure 2: Plots for comparison of Rényi (orange) and Tsallis (blue) distribution with the regular Shannon entropy based distribution (grey) as reference. Noticeably the properties of the Rényi and Tsallis distribution are qualitatively mirrored around $q = 1$.

Note that in all distributions derived from generalization of the Shannon entropy, the familiar Boltzmann exponential behavior is replaced by a power law. Regarding our starting point to incorporate a higher complexity into the considered biological systems than equilibrium thermodynamics suggests, this result is exactly what we longed for [CSN09]. For the Tsallis distribution in Equation (2.46) we find for $q > 1$ that unlikely micro-states are suppressed stronger than likely states narrowing the distribution. This is the case for high evolutionary selection pressure taking effect on the organism. Under these circumstances minor changes of the organism are the appropriate tool to tackle this extraneous influence and to find a surviving sequence analogous to optimization without leaving the vicinity of the local optimum. For $q < 1$, however, likely micro-states are suppressed compared to unlikely states, which is the case for low selection pressure. Then the organism is able to develop alternative strategies with high volatility to ensure its survival. For the Rényi distribution in Equation (2.32) the bahaviour is vice versa as can be seen in Figure 2. Thus, the framework of power law PDs allows us to quantify the evolutionary impact on biological organisms. A question that can be posed is for example: "Is the strong selection pressure induced by intensive drug treatment of HIV protease represented by a Tsallis factor $q > 1$ in the corresponding MSA?"

2.4 Gauge Invariance

In all generalized distributions the Hamiltonian of the system occurs in relation to its expected value. In fact, we could also write the classical Boltzmann factor in the same manner, but due to the additivity law for the arguments of the exponential function, we can reduce any term added to the Hamiltonian with the partition function. We call this *gauge invariance with respect to spectral translations*, which is physically desirable because it implicates the zero level of energy to be arbitrary. In the case of power law distributions this reduction is not possible anymore, but of course we want to maintain this property. The above derived

distributions indeed do so. Consider an arbitrary Hamiltonian and its expected value

$$\sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma}) H(\boldsymbol{\sigma}) = \langle H \rangle \,. \tag{2.48}$$

When adding an arbitrary constant we find

$$\sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma})(H(\boldsymbol{\sigma}) + c) = \langle H \rangle + c \tag{2.49}$$

and the difference $H(\boldsymbol{\sigma}) - \langle H \rangle$ remains the same. This also holds for the q-expected value introduced in Equation (2.36). In addition this form allows for a shift of terms from fields to couplings and vice versa, since the Hamiltonian is the sum of both. This implicates a gauge invariance, because different Hamiltonians lead to the same PD. If not indicated otherwise, throughout this work we will use the *zero sum gauge*, where the sum of all couplings and the sum of all fields along both, rows and columns, is zero. Given a set $\{\hat{\boldsymbol{h}}', \hat{\boldsymbol{J}}'\}$ in arbitrary gauge the corresponding set of $\{\hat{\boldsymbol{h}}, \hat{\boldsymbol{J}}\}$ in zero sum gauge can be obtained by [EHA14]

$$J_{ij}(k,l) = J'_{ij}(k,l) - J'_{ij}(:,l) - J'_{ij}(k,:) + J'_{ij}(:,:) \tag{2.50}$$

$$h_i(k) = h'_i(k) - h'_i(k) + \sum_{j \neq i} \{J'_{ij}(k,:) - J'_{ij}(:,:)\}, \tag{2.51}$$

where ":" is the average over the indicated variable.

By the analytic form of the Rényi and Tsallis distribution functions it is not guaranteed that they are suitable as probability distributions. As soon as the exponent of these functions becomes non integer, complex values for $p(\boldsymbol{\sigma})$ are possible, but undesired. To ensure $p(\boldsymbol{\sigma}) \in [0,1]$, in our numerical investigations we used absolute values of $p(\boldsymbol{\sigma})$. A more detailed discussion about the role of constraints using Tsallis entropy, its implications for thermodynamic relations, and the equivalence of different distributions can be found in [Abe+01; Abe02; Bas04; FMP05b; WS05].

3 Generalized Direct Coupling Analysis

This chapter is dedicated to merging frameworks that have been introduced so far, first the method of direct coupling analysis in Sections 1.1 and 1.2 and secondly the power law probability distributions for the description of discrete one-dimensional networks in Section 2.3. We start out by testing the approximation scheme introduced in Section 1.3 and show that it is not suitable for Tsallis- and Rényi-like probability distributions, since it fails to inhibit a mean field description. This is in fact not at all surprising, since we set out to use these PDs to model non-equilibrium systems with long-range correlations and high complexity, which are by definition not suitable for a mean field description. In a second step, we use the method of pseudo likelihood maximization introduced in Section 1.4 to obtain the equations for developing a generalized DCA algorithm.

3.1 Nonexistence of Mean Field Description

Due to a change of properties of the Gibbs potential when using power law distributions, a naive adoption of the mean-field approximation method introduced in Section 1.3 is not possible for power law PDs. To show this, a relation of the Gibbs potential for power law distributions and the correlation matrix will be attained. Afterwards we will see how the inverse correlation matrix is connected to the interaction Hamiltonian $H_I(\boldsymbol{\sigma})$. In the case of Boltzmann-Gibbs Thermodynamics the following relations hold for the Helmholtz free energy $F = -\ln Z$

$$\frac{\partial F}{\partial h_i(k)} = p_i(k), \tag{3.1}$$

$$\frac{\partial^2 F}{\partial h_i(k)\partial h_j(l)} = C_{ij}(k,l), \tag{3.2}$$

where C is the correlation matrix as defined in Equation (1.16). This gives rise to the conjugated relations of the Gibbs potential

$$G = -\ln Z + \sum_{i,k} h_i(k)p_i(k), \tag{3.3}$$

which is the Legendre transform of F with respect to the fields $h_i(k)$:

$$\frac{\partial G}{\partial p_i(k)} = h_i(k), \tag{3.4}$$

$$\frac{\partial^2 G}{\partial p_i(k)\partial p_j(l)} = (C^{-1})_{ij}(k,l). \tag{3.5}$$

Here $h_i(k)$ and $p_i(k)$ are the conjugate variables.

3.1.1 Derivatives of the Helmholtz Free Energy

We inspect the relations (3.1) and (3.2) for power law distributions. Let $\boldsymbol{\sigma}$ be a micro-state of a physical system governed by the perturbed Hamiltonian as introduced in Equation (1.11)

$$H_\alpha(\boldsymbol{\sigma}) = -\sum_i h_i(\sigma_i) - \alpha \sum_{i<j} J_{ij}(\sigma_i, \sigma_j) \tag{3.6}$$

$$\equiv H_S(\boldsymbol{\sigma}) + \alpha H_I(\boldsymbol{\sigma}). \tag{3.7}$$

Additionally, we introduce the centralized Hamiltonian

$$H^c_\alpha(\boldsymbol{\sigma}) = H_\alpha(\boldsymbol{\sigma}) - \langle H_\alpha \rangle, \tag{3.8}$$

with the average Hamiltonian constituted of the averages of the above introduced single-site and interaction Hamiltonian respectively

$$\langle H_\alpha \rangle = -\sum_i \sum_k h_i(k) p_i(k) - \alpha \sum_{i<j} \sum_{k,l} J_{ij}(k,l) p_{ij}(k,l) \tag{3.9}$$

$$\equiv \langle H_S \rangle + \alpha \langle H_I \rangle . \tag{3.10}$$

In order to be able to amalgamate the following calculations we generalize the various different power law distributions seen in Section 2.3. Then the partition function reads

$$Z(\alpha) = \sum_{\boldsymbol{\sigma}} \mathcal{F}(H_\alpha^c(\boldsymbol{\sigma}))^\gamma, \tag{3.11}$$

with $\gamma \in \mathbb{R}$ and a linear function $\mathcal{F}(H_\alpha^c(\boldsymbol{\sigma})) = c_0 + c_1 H_\alpha^c(\boldsymbol{\sigma})$, with $c_0 \in \mathbb{R}$, $c_1 \in \mathbb{R} \setminus \{0\}$. For the Tsallis distribution we have $\gamma = 1/c_1 = 1/1-q$ and $c_0 = 1$. For the Rényi distribution we have $\gamma = 1/qc_1 = 1/q-1$ and again $c_0 = 1$. Let $F(\alpha) = -\ln Z(\alpha)$. Detailed calculations are given in Appendix B.2. We find

$$\frac{\partial F(\alpha)}{\partial h_i(k)} = \gamma \sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma}) \frac{\delta_{\sigma_i k} - f_i(k)}{\zeta + H_\alpha^c(\boldsymbol{\sigma})} \equiv \pi_i(k), \tag{3.12}$$

where we call $\pi_i(k)$ the *energy weighted probability* to find spin k at position i, and $\zeta = c_o/c_1$ is a real valued constant. The obtained result is very much related to the original Boltzmann-Gibbs result. In the limit of $\mathcal{F}^\gamma$ being replaced by the exponential function the original probability $p_i(k)$ is recovered. For further simplification this term can be linearized in the centralized Hamiltonian, which is by definition distributed around 0:

$$\pi_i(k) \approx \frac{\gamma}{\zeta^2} \left\{ f_i(k) \langle H_\alpha \rangle - \left\langle \delta_{\sigma_i k} H_\alpha \right\rangle \right\}, \tag{3.13}$$

where $\left\langle \delta_{\sigma_i k} H_\alpha \right\rangle$ is the expected energy of all sequences in which spin k sits at residue i. The linearized terms are calculated in Appendix B.3. Furthermore we calculate

$$\frac{\partial^2 F}{\partial h_i(k) \partial h_j(l)} = \gamma(\gamma - 1) \sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma}) \frac{(\delta_{\sigma_i k} - f_i(k))(\delta_{\sigma_j l} - f_j(l))}{(\zeta + H_\alpha^c(\boldsymbol{\sigma}))^2} - \pi_i(k)\pi_j(l) \tag{3.14}$$

$$\equiv \pi_{ij}(k,l) - \pi_i(k)\pi_j(l) \equiv K_{ij}(k,l), \tag{3.15}$$

with the *energy weighted correlation matrix* K. Upon linearization in $H_\alpha^c(\boldsymbol{\sigma})$ we find

$$\begin{aligned} \pi_{ij}(k,l) \approx & \frac{\gamma(\gamma-1)}{\zeta^2} \{ f_{ij}(k,l) \left(1 - \frac{2}{\zeta} \langle H_\alpha \rangle \right) - f_i(k) f_j(l) \left(1 + \frac{4}{\zeta} \langle H_\alpha \rangle \right) \\ & + \frac{2}{\zeta} \left\{ \left\langle \delta_{\sigma_i k} \delta_{\sigma_j l} H_\alpha \right\rangle + \left\langle \delta_{\sigma_i k} H_\alpha \right\rangle f_j(l) + \left\langle \delta_{\sigma_j l} H_\alpha \right\rangle f_i(k) \right\} \}. \end{aligned} \tag{3.16}$$

We know from the properties of the Legendre transformation that given the conjugated variable $\pi_i(k)$, it holds that

$$\frac{\partial^2 G}{\partial \pi_i(k) \partial \pi_j(l)} = (K^{-1})_{ij}(k,l), \tag{3.17}$$

but this is not very helpful since neither $\pi_i(k)$ nor $K_{ij}(k,l)$ are accessible from the MSA. This is due to the unfortunate coupling between quantities accessible via an MSA like $\left\langle \delta_{\sigma_i k} \right\rangle = f_i(k)$ and desired quantities like $\langle H \rangle$. For this problem we could find any workaround.

3.1.2 Derivatives of the Gibbs Potential

We derive the derivatives of G with respect to the accessible variables $p_i(k)$ in order to find a relation like

$$(C^{-1})_{ij}(k,l) = -J_{ij}(k,l) \tag{3.18}$$

for $i \neq j$, as holds in the Boltzmann-Gibbs case. Since contact prediction depends on a threshold of the Frobenius norm [GVL12] of the $Q \times Q$ matrix J_{ij}, a similar relation would be a promising requirement for the method introduced in Section 1.3 if the correlation matrix was accessible via the MSA. However, in this section we show, that a linearization in α and $H^c_\alpha(\boldsymbol{\sigma})$ also gives no further insight, as the derivative of the first term in the Taylor expansion vanishes. Using the same definitions as above we calculate

$$\frac{\mathrm{d}G(\alpha)}{\mathrm{d}\alpha} = -\gamma \sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma}) \frac{H_l(\boldsymbol{\sigma}) - \langle H_l \rangle}{\zeta + H^c_\alpha(\boldsymbol{\sigma})}, \tag{3.19}$$

with the real valued constant $\zeta = c_o/c_1$. Due to the non-linear relation of the probability and the Hamiltonian in this equation a differentiation with respect to $p_i(k)$ and $p_j(l)$ is impossible. Nonetheless, we can again linearize in $H^c_\alpha(\boldsymbol{\sigma})$:

$$\left.\frac{\mathrm{d}G(\alpha)}{\mathrm{d}\alpha}\right|_{\alpha=0} \approx \frac{\gamma}{\zeta^2} \left\{ \langle H_S H_l \rangle - \langle H_S \rangle \langle H_l \rangle \right\}. \tag{3.20}$$

Now differentiation is possible (see Appendix B.4) and we find for $\alpha = 0$

$$\frac{\partial^2}{\partial p_i(k) \partial p_j(l)} \langle H_S H_l \rangle = \frac{\partial^2}{\partial p_i(k) \partial p_j(l)} \langle H_S \rangle \langle H_l \rangle \tag{3.21}$$

We have to conclude that a naive adoption of the mean field approximation does not hold for power law probability distributions.

3.2 Pseudo Likelihood Maximization

Since a simple mean field expansion to solve the inverse Ising problem is not possible in the case of a power law distribution using Tsallis or Rényi entropy, we employ the method of pseudo likelihood maximization as it was introduced for the Boltzmann-Gibbs distribution in the classical DCA in Section 1.4. This method has recently attracted more and more attention throughout other scientific branches. Let us list as examples spin glasses [CB14] and a paper with general approaches [DRT14]. In this case the conditional probabilities of finding a certain spin at a specific position in the sequence, given the rest of the sequence is known, have to be calculated. Therefore let us introduce the energy of such a case given by the Hamiltonian

$$\begin{aligned} H(k|\boldsymbol{\sigma}_{\setminus r}) = &-h_r(k) - \sum_{i<r} J_{ir}(\sigma_i, k) - \sum_{j>r} J_{rj}(k, \sigma_j) \\ &- \sum_{i \neq r} h_i(\sigma_i) - \sum_{\substack{i<j \\ i,j \neq r}} J_{ij}(\sigma_i, \sigma_j), \end{aligned} \tag{3.22}$$

where spin k resides at position r and the rest of the sequence is called $\boldsymbol{\sigma}_{\setminus r}$. The associated centralized Hamiltonian is then called

$$H^c(k|\boldsymbol{\sigma}_{\setminus r}) = H(k|\boldsymbol{\sigma}_{\setminus r}) - \langle H \rangle, \tag{3.23}$$

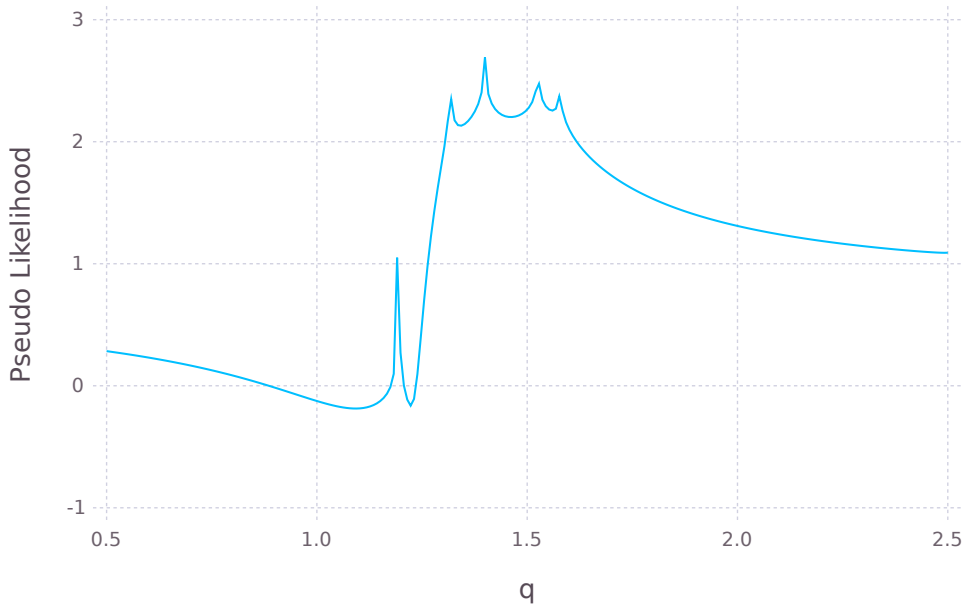

Figure 3: Plot of the negative logarithmic pseudo likelihood for different q and a fixed parameter set $\{\hat{\boldsymbol{h}}, \hat{\boldsymbol{J}}\}$. An MSA has been generated for a $Q = 4$, $L = 4$ toy model via Metropolis' classical algorithm. The original Boltzmann DCA method was then used to find the best fit of fields and couplings. Due to the non-convexity, appropriate initial values for the local optimization algorithm have to be found.

where the expected value $\langle H \rangle$ is the weighted sum over the marginals as usual. With the Tsallis probability distribution and its partition function

$$p(\boldsymbol{\sigma}) = \frac{1}{Z}(1 - (1-q)H^c(\boldsymbol{\sigma}))^{\frac{1}{1-q}}, \tag{3.24}$$

$$Z = \sum_{\boldsymbol{\sigma}}(1 - (1-q)H^c(\boldsymbol{\sigma}))^{\frac{1}{1-q}}, \tag{3.25}$$

we find the conditional probability to be

$$p(k|\boldsymbol{\sigma}_{\setminus r}) = \frac{Q_{k_r}}{\sum_m Q_{m_r}} \equiv \frac{(1 - (1-q)H^c(k|\boldsymbol{\sigma}_{\setminus r}))^{\frac{1}{1-q}}}{\sum_m (1 - (1-q)H^c(m|\boldsymbol{\sigma}_{\setminus r}))^{\frac{1}{1-q}}}, \tag{3.26}$$

where we introduced a short form of the numerator of the Tsallis probability Q_{k_r} when spin k is at position r and the rest of the sequence $\boldsymbol{\sigma}_{\setminus r}$ is known. A similar expression as (3.26) can be deduced using the Rényi distribution as found in Section 2.3.1

$$p(k|\boldsymbol{\sigma}_{\setminus r})_{\mathrm{R}} = \frac{(1 - \frac{q-1}{q}H^c(k|\boldsymbol{\sigma}_{\setminus r}))^{\frac{1}{q-1}}}{\sum_m (1 - \frac{q-1}{q}H^c(m|\boldsymbol{\sigma}_{\setminus r}))^{\frac{1}{q-1}}}. \tag{3.27}$$

Since both distributions are qualitatively the same only for different values for q, we restrict ourselves from now on to only examine the Tsallis case. Nevertheless, the following calculations work completely analogously for both distributions.

Note that in contrast to the Boltzmann-Gibbs conditional probability in Equation (1.29), where the summation has to be performed only over fields and couplings of residue r, in the Tsallis case the energy of the entire sequence has to be calculated. Thus, a residue-wise evaluation of the likelihood function becomes impossible. Such a procedure has been

introduced in [EHA14] and called *asymmetric pseudo likelihood maximization*, due to the fact that it will in general result in different estimates for $J_{ij}(k,l)$ and $J_{ji}(l,k)$, which then have to be combined, since analytically they are supposed to be the same. However, the arising subproblem has much less free parameters to be optimized and it allows a trivial parallelization of the process over the sequence length L, largely increasing the algorithm's performance. Since the conditional probabilities in the Tsallis distribution depend on the entire sequence, this will lead to an increase in calculation time of order L compared to the classical symmetric pseudo likelihood maximization DCA algorithm (plmDCA), where $\mathcal{O}(QBL^2)$ and a single optimization for all parameters has to be performed. The conditional probabilities for every residue yield the negative logarithmic pseudo likelihood function as the sum over all sequences in an MSA

$$l_{\text{pseudo}} = -\frac{1}{B_{\text{eff}}} \sum_{b=1}^{B} \frac{1}{m_b} \sum_{r=1}^{L} \left\{ \ln Q_{\sigma_r}^{(b)} - \ln \sum_k Q_{k_r}^{(b)} \right\} \tag{3.28}$$

In order to efficiently solve the posed optimization problem we also give an analytical form of the derivatives of l_{pesudo} with respect to the fields, couplings, and Tsallis parameter, which values are to be optimized. The derivative with respect to $h_i(k)$:

$$\begin{aligned}\frac{\partial l_{\text{pseudo}}}{\partial h_i(k)} = -\frac{1}{B_{\text{eff}}} \sum_{b=1}^{B} \frac{1}{m_b} \{ \sum_{r=1}^{L} \left\{ Q_{\sigma_r}^{(b)^{q-1}} - \frac{1}{\sum_n Q_{n_r}^{(b)}} \sum_m Q_{m_r}^{(b)^q} \right\} (\delta_{\sigma_i k} - f_i(k)) \\ - \frac{1}{\sum_n Q_{n_i}^{(b)}} \sum_m Q_{m_i}^{(b)^q} (\delta_{mk} - \delta_{\sigma_i k}) \}.\end{aligned} \tag{3.29}$$

The derivative with respect to $J_{ij}(k,l)$:

$$\begin{aligned}\frac{\partial l_{\text{pseudo}}}{\partial J_{ij}(k,l)} = -\frac{1}{B_{\text{eff}}} \sum_{b=1}^{B} \frac{1}{m_b} \{ \sum_{r=1}^{L} \left\{ Q_{\sigma_r}^{(b)^{q-1}} - \frac{1}{\sum_n Q_{n_r}^{(b)}} \sum_m Q_{m_r}^{(b)^q} \right\} (\delta_{\sigma_i k}\delta_{\sigma_j l} - f_{ij}(k,l)) \\ - \frac{1}{\sum_n Q_{n_i}^{(b)}} \sum_m Q_{m_i}^{(b)^q} (\delta_{mk} - \delta_{\sigma_i k})\delta_{\sigma_j l} \\ - \frac{1}{\sum_n Q_{n_j}^{(b)}} \sum_m Q_{m_j}^{(b)^q} (\delta_{ml} - \delta_{\sigma_j l})\delta_{\sigma_i k} \}.\end{aligned} \tag{3.30}$$

The derivative with respect to q:

$$\begin{aligned}\frac{\partial l_{\text{pseudo}}}{\partial q} = -\frac{1}{B_{\text{eff}}} \sum_{b=1}^{B} \frac{1}{m_b} \sum_{r=1}^{L} \frac{1}{1-q} \{ \ln Q_{\sigma_r}^{(b)} + Q_{\sigma_r}^{(b)^{q-1}} H^c(\boldsymbol{\sigma}^{(b)}) \\ - \frac{1}{\sum_n Q_{n_r}^{(b)}} \sum_m Q_{m_r}^{(b)} \left\{ \ln Q_{m_r}^{(b)} + Q_{m_r}^{(b)^{q-1}} H^c(m|\boldsymbol{\sigma}_{\setminus r}^{(b)}) \right\} \}.\end{aligned} \tag{3.31}$$

It is not obvious from the analytical form of the likelihood function, whether it is convex, as it is the case for such a function based on the Boltzmann-Gibbs distribution. Throughout our research we tested the convexity numerically and found it not to be the case. In Figure 3, this is shown for the variation of q and in Figure 4 for variation of a local field parameter. To find the global minimum of a non-convex function, a local optimizer does not suffice. Yet, testing of a *Multi-Level Single-Linkage* global optimization scheme [KS05], revealed that it is not feasible for the high number of parameters protein sequences inhibit, significantly increasing the computation time. Therefore, appropriate initial values have to be found in order to find the global minimum with a local optimization scheme. For this, we consult the classical DCA algorithm based on the Boltzmann-Gibbs distribution.

When using the logarithm of the pseudo likelihood, as we do, the generalized form $\ln_q$ that was introduced in Equation (2.9), also appears to be a possible ansatz instead of the regular logarithm. We investigated this approach, which has additionally been done in [Yam02], and rejected it due to the unfeasible analytic form it leads to.

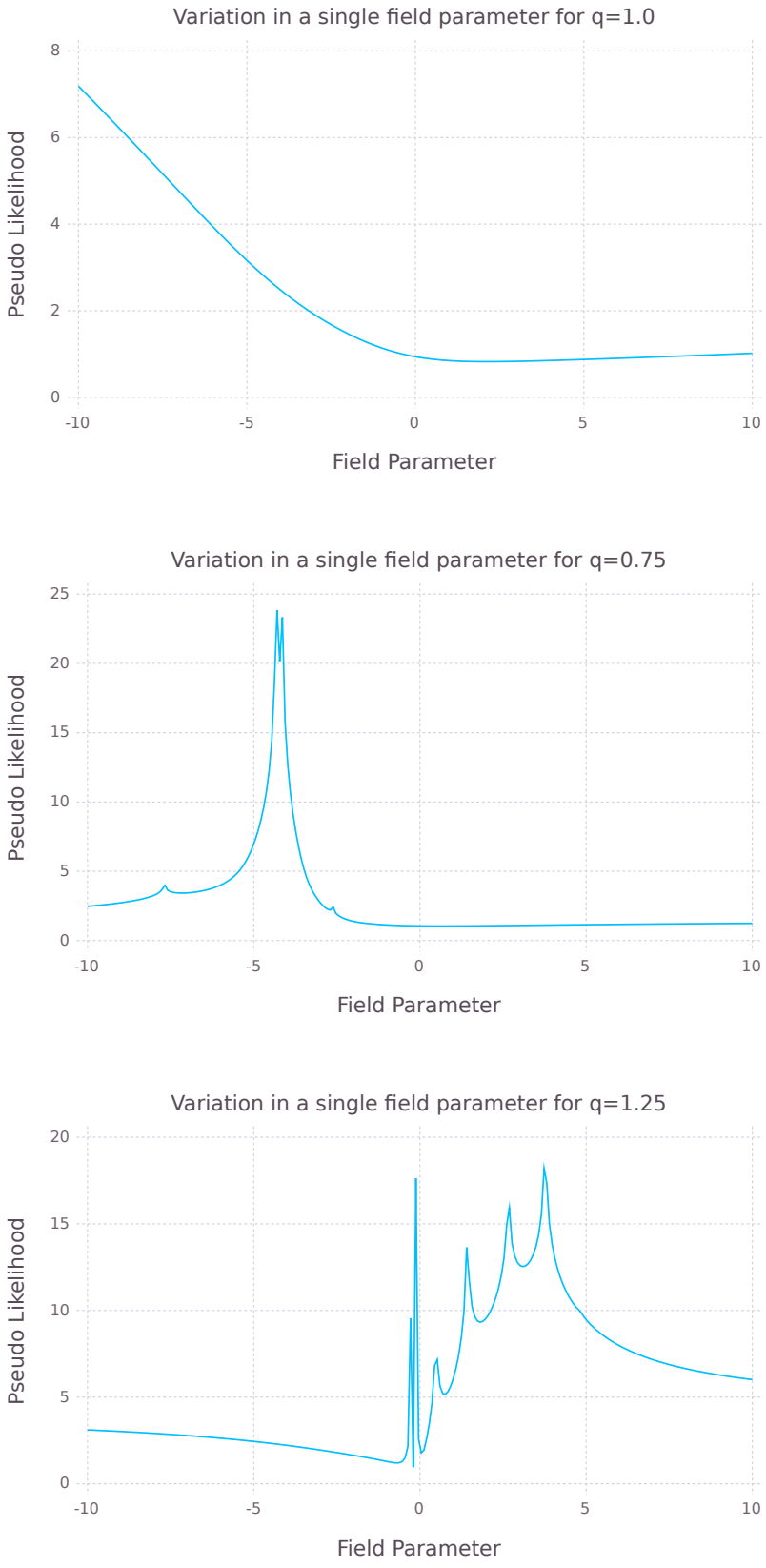

Figure 4: While the pseudo likelihood function is convex in the fields and couplings for Boltzmann-Gibbs distributed sequences, this is not the case for the Tsallis distribution. To simplify the presentation, only variation of a single field parameter of the multi-dimensional function is plotted here.

4 Implementation and Results

After development of an analytical theory of a *generalized direct coupling analysis* (GDCA) we now investigate the numerical approach and the implementation of an algorithm, which can generate a contact map from a given multiple sequence alignment. Our working hypothesis is that in the classical DCA model, i.e., with Shannon entropy, probable sequences are not correctly represented in biological data, due to varying evolutionary time scales in protein production and sampling biases. This will lead to higher likelihood values for q-values deviating from the classical Shannon entropy ($q = 1$). Higher likelihood values indicate that the theoretical protein model used in GDCA is the appropriate tool to satisfy the data. Synthetic data will be generated using Metropolis' algorithm [Met+53] and an optimization procedure overcoming the non-convexity issue will be introduced. Our preferred toy model throughout this study is a four residue system with a four letter alphabet causing a partition function of size 256 and 112 independent field and coupling parameters. This model can be both, solved exactly and optimized, within minutes on a single core[2], while still inhibiting decent complexity. We also test the developed algorithm on a real protein, namely Interleukin8 (CXCL6) from the Chemokine family (see Figure 1). This is a 61 residue protein with the full 21 letter alphabet leading to a partition function of size $> 4.5 \times 10^{80}$ and 807,030 parameters, which have to be optimized. On a single core a GDCA optimization for a fixed q takes approximately 20 hours.

4.1 Algorithm Development

With Metropolis' algorithm it is possible to generate sequence data according to the Boltzmann-Gibbs distribution. After omitting enough sequences such that thermodynamic equilibrium is ensured, a randomly altered sequence $\boldsymbol{\sigma}'$ will be preferred over the original sequence $\boldsymbol{\sigma}$ with probability

$$p^{\mathrm{B}}_{\mathrm{Metro}} = \min\{1, \exp\big\{H(\boldsymbol{\sigma}) - H(\boldsymbol{\sigma}')\big\}\}. \tag{4.1}$$

This acceptance condition can easily be generalized for the Tsallis distribution

$$p^{\mathrm{T}}_{\mathrm{Metro}} = \min\{1, \left(\frac{1-(1-q)\left\{H(\boldsymbol{\sigma}') - \langle H\rangle\right\}}{1-(1-q)\left\{H(\boldsymbol{\sigma}) - \langle H\rangle\right\}}\right)^{\frac{1}{1-q}}\} \tag{4.2}$$

or Rényi distribution

$$p^{\mathrm{T}}_{\mathrm{Metro}} = \min\{1, \left(\frac{1-\frac{q-1}{q}\left\{H(\boldsymbol{\sigma}') - \langle H\rangle\right\}}{1-\frac{q-1}{q}\left\{H(\boldsymbol{\sigma}) - \langle H\rangle\right\}}\right)^{\frac{1}{q-1}}\}. \tag{4.3}$$

The performance of the method in matching the original partition function for various sample sizes is shown in Figure 5. It is now possible to create toy model MSAs, where the field and coupling parameters are known for various values of q to benchmark the developed GDCA algorithms.

The pseudo likelihood function based on the Tsallis distribution is not convex in the variation of field and coupling parameters as well as in q. This can be seen in Figures 4 and 3, respectively. Optimizing fields, couplings, and q in the same scheme leads to serious convergence issues. Albeit, it can be done separately, which we call *parameter GDCA for fixed q* (pGDCA) and *Tsallis q GDCA for fixed parameters* (qGDCA), respectively. The method of classical DCA starts its optimization process with zero fields and couplings as initial values. This information was obtained by rebuilding and comparing the result with

[2] Intel®Core™2 Quad CPU Q9550 @ 2.83GHz

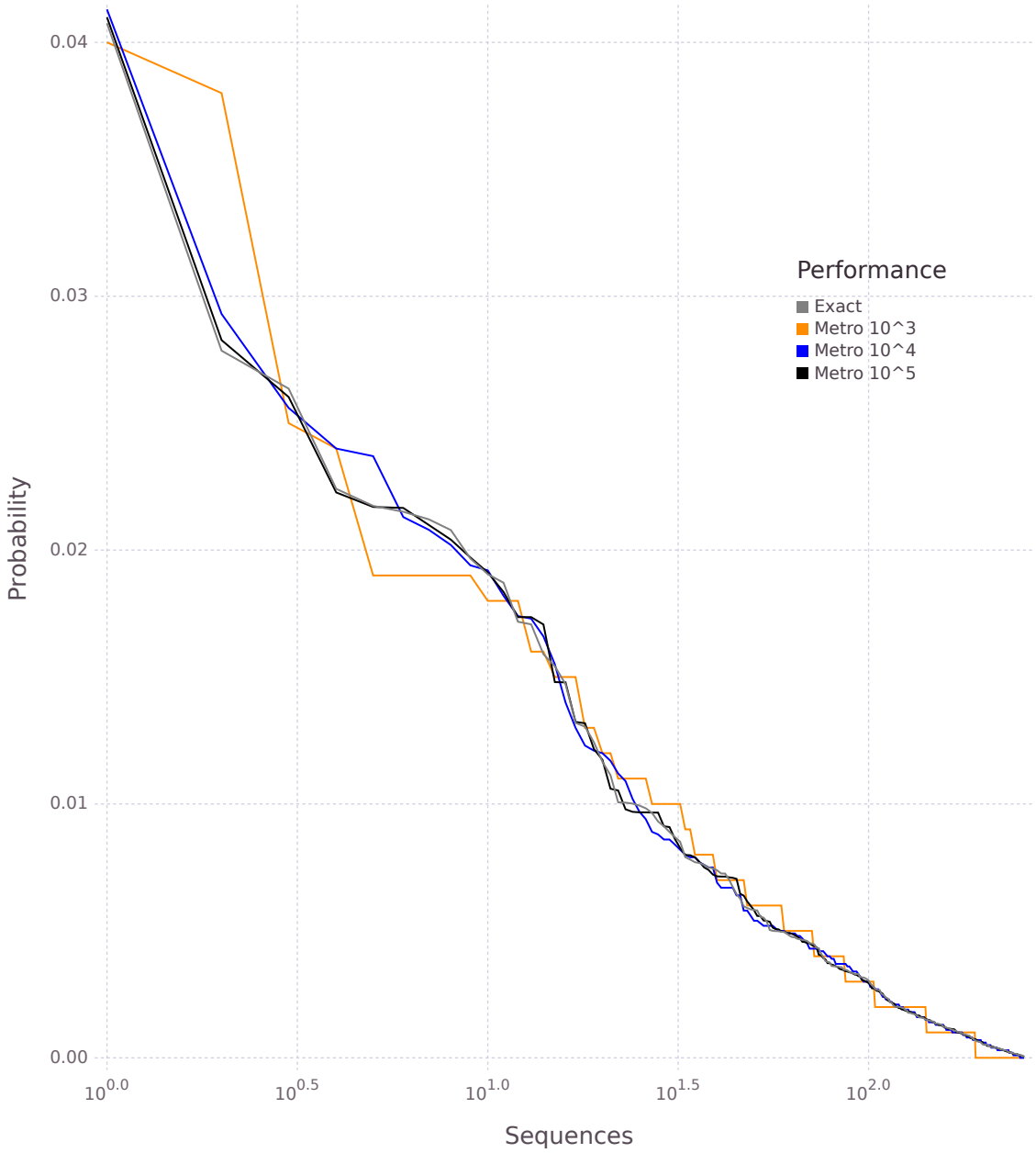

Figure 5: The performance of the Metropolis algorithm for the Boltzmann distribution increases with the number of generated sequences. Plotted are the various sample sizes for an artificial example of a 4 letter alphabet in 4 residues. Note that for large sample sizes (black) the distribution is practically matching the exact calculation (grey). The Tsallis and Rényi distribution also meet these findings for relevant q values ± 0.3 around 1. For other q, the acceptance criterion has to be tuned in order to avoid freezing of a single random sequence.

the algorithm for symmetric plmDCA as proposed in [EHA14] and published on GitHub [Git] in the Julia programming language [Jul]. From there it is straight forward to generalize the algorithm to the equations derived in Section 3.2. While zero fields and couplings as initial values still lead to convergence of the local optimizer, it may not be a global optimum and variation of the initial values leads to different optima. However, a global optimization scheme has shown to be not feasible because of extremely long convergence times, even for small toy models with only ~ 10 parameters. We therefore chose to use the Limited-memory Broyden-Fletcher-Goldfarb-Shanno optimization scheme (LBFGS) [Noc80; LN89], which is very powerful regarding convergence time and memory efficiency. It is a local optimizer, which uses committed analytical derivatives, hence the necessity of Equations (3.29) - (3.31). In order to ensure convergence to the actual global optimum we implemented an iterative procedure as follows. As a first step we used our implementation of the classical DCA method to acquire initial values for the pGDCA algorithm. This is sensible under the assumption that the Boltzmann-Gibbs distribution holds as a first estimate of the underlying model and that the proper q value is in the vicinity of 1. The results prove this assumption to be well met. Then the area around $q = 1$ is tested step by step. We chose $\Delta q = 0.01$ as step size in most cases, being a practical trade-off between resolution and computation time. The results for the field and couplings were used as initial values for the next step. It is not always possible to sweep the whole area, for example between $q = 0$ and $q = 2$, as we intended to, because convergence of the algorithm can not be guaranteed. Nonetheless, in all our calculations an unambiguous optimum could be identified, which is why for large scale calculations, we stopped the iterative procedure as soon as the pseudo likelihood did not exceed its predecessor. Yet, we can not guarantee the optimum to be global, although it clearly is an improvement compared to the classical model. We call this method *Tsallis direct coupling analysis* (TDCA).
An alternative procedure, which is not as accurate, but significantly quicker due to much less optimization steps was used to generate vivid, but not necessarily exact plots as seen in Figure 6, revealing the theoretical potential of our idea. The first step is the same as in TDCA. For the attained parameter set s_1 the optimal q value is not necessarily one, however, it can be calculated using qGDCA, the q optimization for fixed fields and couplings. Then the pGDCA can be performed for the found q, with s_1 as initial values. This leads to a new, possibly better, parameter set s_2. We have done that exemplary for sequences of the mentioned toy model generated with a single randomly chosen parameter set. Figure 6 exhibits the negative logarithmic pseudo likelihood for varying q with the parameter set s_1 and for s_2, both for two exemplary different sample sizes. More sample sizes were tested and we observed increasing overlap of the two plots, the more sequences the MSA contained. In every calculation there is a minimum of l_{pseudo} for $q \neq 1$. This suggests that the Tsallis distribution is the appropriate model for sequence occurrence, especially for practical MSA sizes much smaller than the partition function. For different parameter sets used to generate the sequences, the form of the plots may vary, for example the distribution may qualitatively be mirrored around $q = 1$.

4.2 Numerical Results

Figure 6 suggests that the classical DCA method is suitable for large sample sizes, and in such cases, the data is not in any way better represented by the Tsallis distribution. However, TDCA can still improve the prediction fidelity, as illustrated Figure 9, which will be discussed below. Additionally, for smaller sample sizes, as they are common for proteins, Tsallis DCA poses a serious challenge to the known method, as l_{pseudo} is underbid. Starting with these results, which reveal that the q value of an MSA may be depending on its size, we used the TDCA method to generate data for the $Q^L = 4^4$ toy model introduced at the beginning

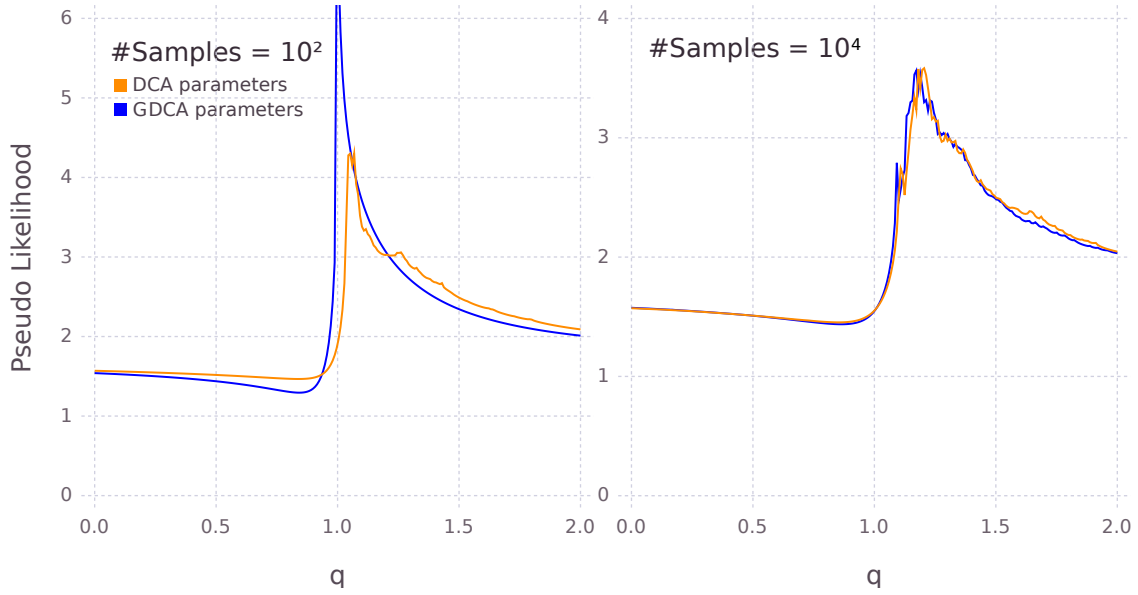

Figure 6: For larger but finite sample size the classical Boltzmann DCA increases in overlap with the Tsallis DCA. The optimal parameter set after optimization with the classical DCA algorithm (orange) and with the pGDCA algorithm (blue) are plotted for varying q. We find the pseudo likelihood to be lower for $q \neq 1$ compared to $q = 1$.

of this chapter. Two main sequence generation methods were used, both with the classical Metropolis method, i.e. for $q = 1$. First, the parameter set $\{\hat{\boldsymbol{h}}, \hat{\boldsymbol{J}}\}$ was randomly chosen from a uniform distribution and then fixed. Generating sequences from this set of fields and couplings is apt to gather information about sequencing biases depending on the sample size. As a second method, for each generated MSA we used new randomly chosen parameter sets. This method is suited for general statements about the coupling analysis under investigation. Both methods were tested for both, performance in the sense of reproducing the input parameters and performance in the sense of pseudo likelihood maximization (to be precise minimization of the normalized negative logarithmic pseudo likelihood). For both methods a thousand data points were recorded each for seven representative sample sizes ranging from 0.2 to 50 times the size of the number of terms in the partition function Q^L. The sample sizes and the key data about q for the optimal pseudo likelihood value are reprinted in Appendix A in Table 1 and visualized in Figure 7. We find that only for MSAs of size much larger than the partition function ($\gtrsim 20 \times Q^L$) an equilibrium of $q = 1$ will be reached. This is as expected for sufficiently large samples, since the sequences are generated for that value of q. For sample sizes around Q^L up $10 \times Q^L$ the data shows a high volatility around 1 for both, fixed and random parameters. This indicates clearly that q is dependent on the specific sample under investigation and not on the parameter set for these sample sizes, although the sequences are generated from the Boltzmann-Gibbs distribution. For MSAs smaller than the size of the partition function, as it always is the case for real biological data, we observe a significant bias towards $q < 1$. Note that for random parameter sets the average q is still well under 1, although there are a few outliers greater than 1. In general, we can say that for random parameter sets the volatility of the results is greater than for a fixed parameter set, whereas the average q values follow the same pattern. Still, the improvement in pseudo likelihood is very small as shown in Figure 8, where l_{pseudo} (in the plots also referred to as PL) is plotted for the optimal found q and for $q = 1$. Notable differences between the two sequence generation methods become obvious in these plots: The distribution of l_{pseudo} values is rather wide for randomly chosen parameter sets, whereas for the same set it has a strong peak around $l_{\text{pseudo}} \approx 3.96$ as indicated by the color in Figure 8. Figure 9 visualizes

the results of the performance of the TDCA algorithm with regard to reproduction of the original fields and couplings the samples were generated with $\{\hat{\boldsymbol{h}}, \hat{\boldsymbol{J}}\}$. Then we call the result of TDCA $\{\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\zeta}}\}$ and we introduce the fidelity measure

$$f = f_h + f_J, \tag{4.4}$$

$$f_h = \frac{1}{QL} \sum_i \sum_k \left| \frac{\eta_i(k) - h_i(k)}{h_i(k)} \right|, \tag{4.5}$$

$$f_J = \frac{2}{Q^2 L(L-1)} \sum_{i<j} \sum_{k,l} \left| \frac{\zeta_{ij}(k,l) - J_{ij}(k,l)}{J_{ij}(k,l)} \right|. \tag{4.6}$$

The constant fractions before the sums consider the number of terms in the respective sum, making the fidelity f effectively an average relative deviation of the single parameters. This is a measure that strongly penalizes discrepancies, where the original fields and couplings are small. Other measures are in principle also possible, for example of the form

$$g = g_h + g_J, \tag{4.7}$$

$$g_h = \frac{1}{QL} \sum_i \sum_k \left| \frac{\eta_i(k) - h_i(k)}{\eta_i(k) + h_i(k)} \right|, \tag{4.8}$$

$$g_J = \frac{2}{Q^2 L(L-1)} \sum_{i<j} \sum_{k,l} \left| \frac{\zeta_{ij}(k,l) - J_{ij}(k,l)}{\zeta_{ij}(k,l) + J_{ij}(k,l)} \right|. \tag{4.9}$$

but we preferred f over g, because actual contact prediction depends on the Frobenius norm of calculated parameters $\{\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\zeta}}\}$, which takes into account absolute values of the individual matrix entries. Thus, errors, where the fields are supposed to be small, have a strong impact on DCA results. On the other hand, if one of the two values of $\{\hat{\boldsymbol{h}}, \hat{\boldsymbol{J}}\}$ and $\{\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\zeta}}\}$ is already very large, even a high deviation between the two does not reflect appropriately in the case of g, justifying our choice of fidelity measure. With TDCA we outperform the classical DCA clearly for sample sizes $\gtrsim 3 \times Q^L$. For a model of the size of our toy model the reweighting in the original DCA algorithm does not come into effect with the standard threshold implemented. Therefore no comparison with an artificial reweighting was exerted. For smaller MSAs TDCA tends to overestimate the parameters.

Testing TDCA on Interleukin8 (CXCL8) did not lead to new insights apart from ensuring its operability. The optimal q we found was exactly 1 for both, step sizes of $\Delta q = 0.1$ and $\Delta q = 0.01$. A plot can be found in Appendix A. The alignment was taken from the PsiCov database [Jon+12].

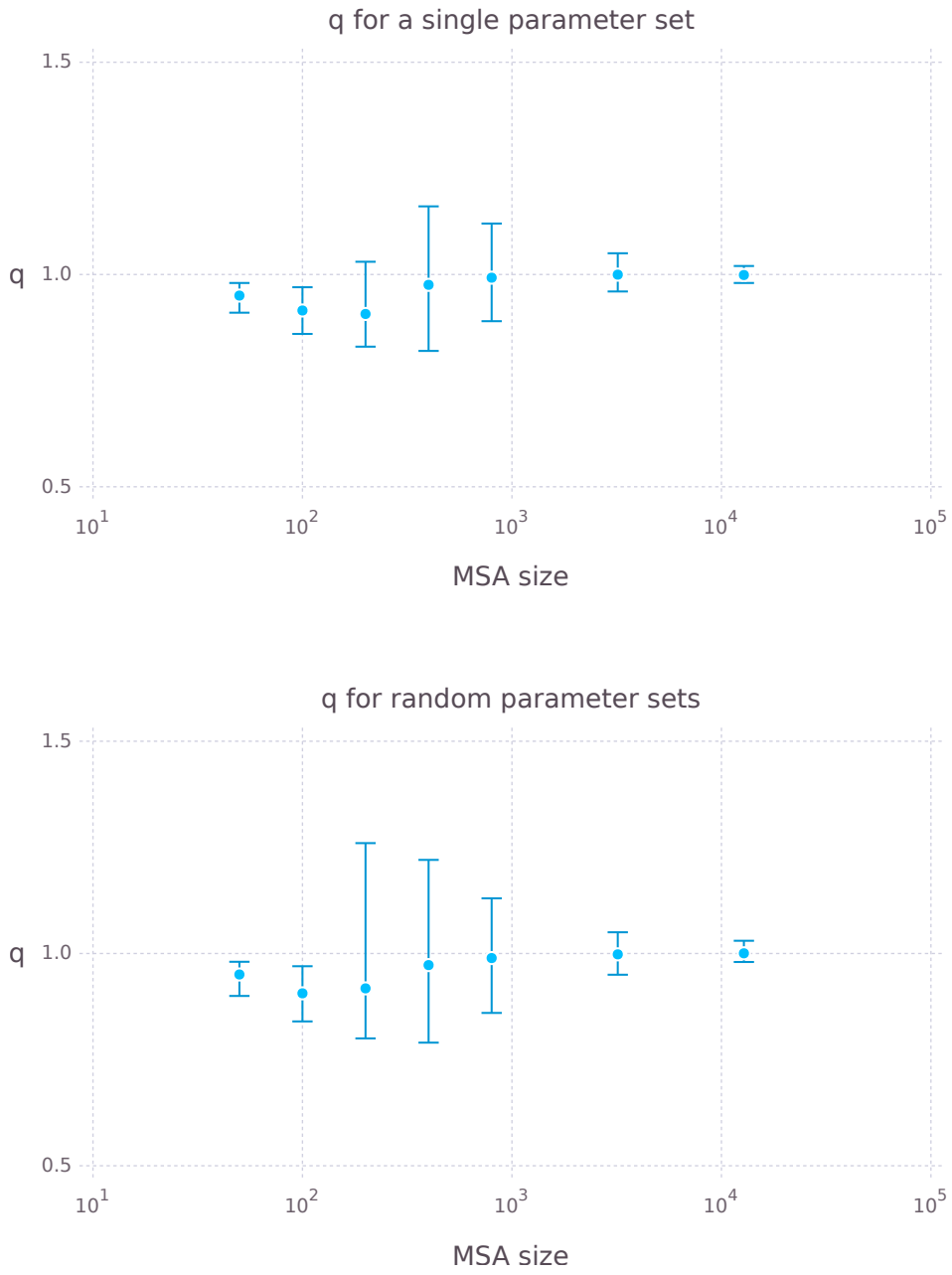

Figure 7: The summarized data of a thousand optimization procedures, each for varying sample sizes generated with $q = 1$, show a significant bias towards $q < 1$ when the sample size is below the number of sequences in the partition function. In the shown case this number was 256 for a toy model with a four letter alphabet at four residues. The upper plot shows the mean and 95% quantiles of multiple samples generated from a single randomly chosen parameter set, whereas in the lower plot each sample is generated by a different randomly chosen parameter set. The numerical results can also be seen in Table 1.

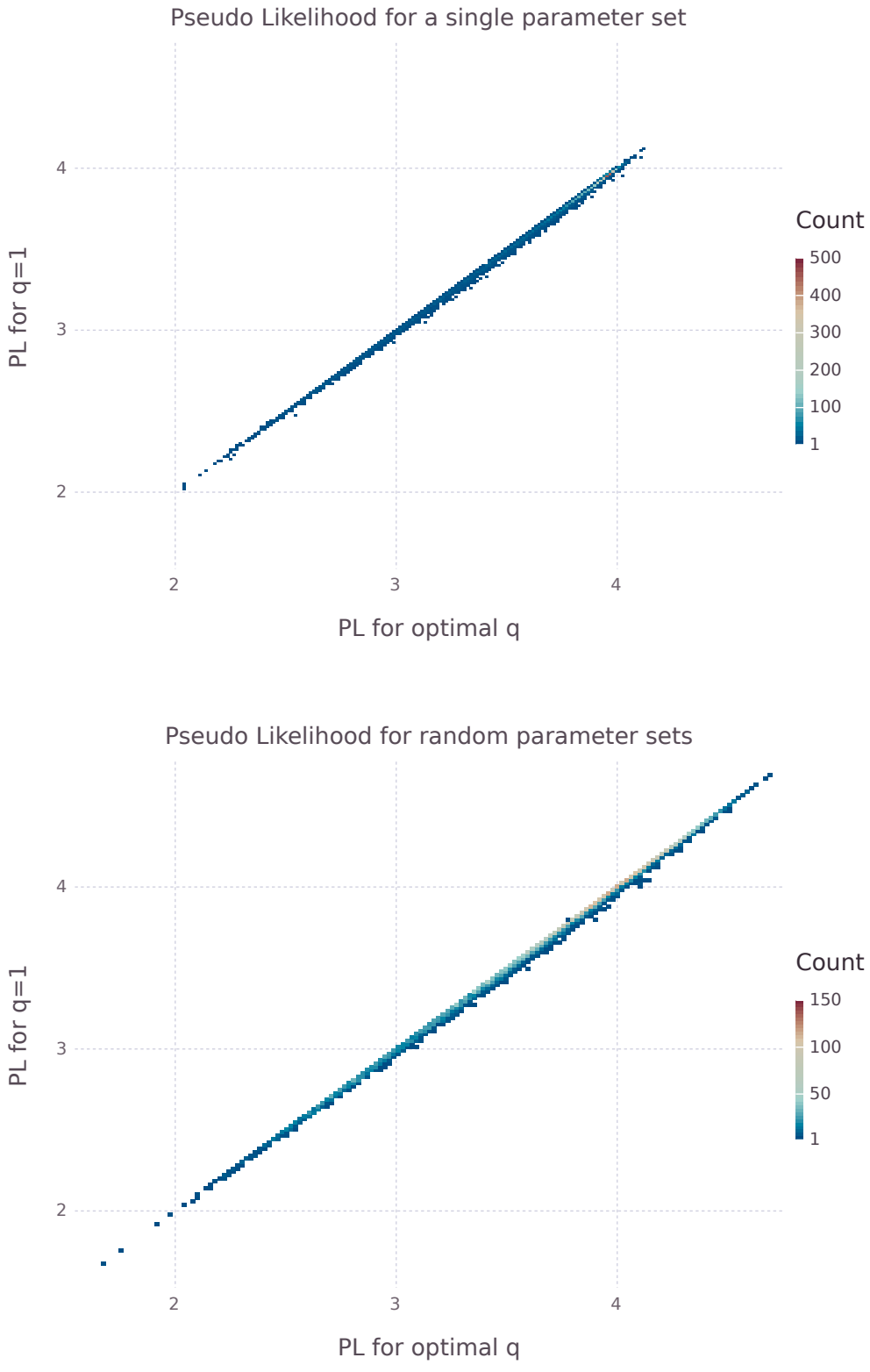

Figure 8: The TDCA algorithm improves the pseudo likelihood compared to the DCA algorithm from [EHA14]. This is a histogram plot of each 1000 measurements for 7 sample sizes from 0.2 to 50 times the partition function size. There is no significant difference in improvement as a function of sample size. The upper plot shows the result for samples taken from one single fixed parameter set inhibiting a major peak at $I_{\text{pseudo}} \approx 3.96$, whereas for the lower plot a new parameter set was diced for each sample showing a widespread range of values.

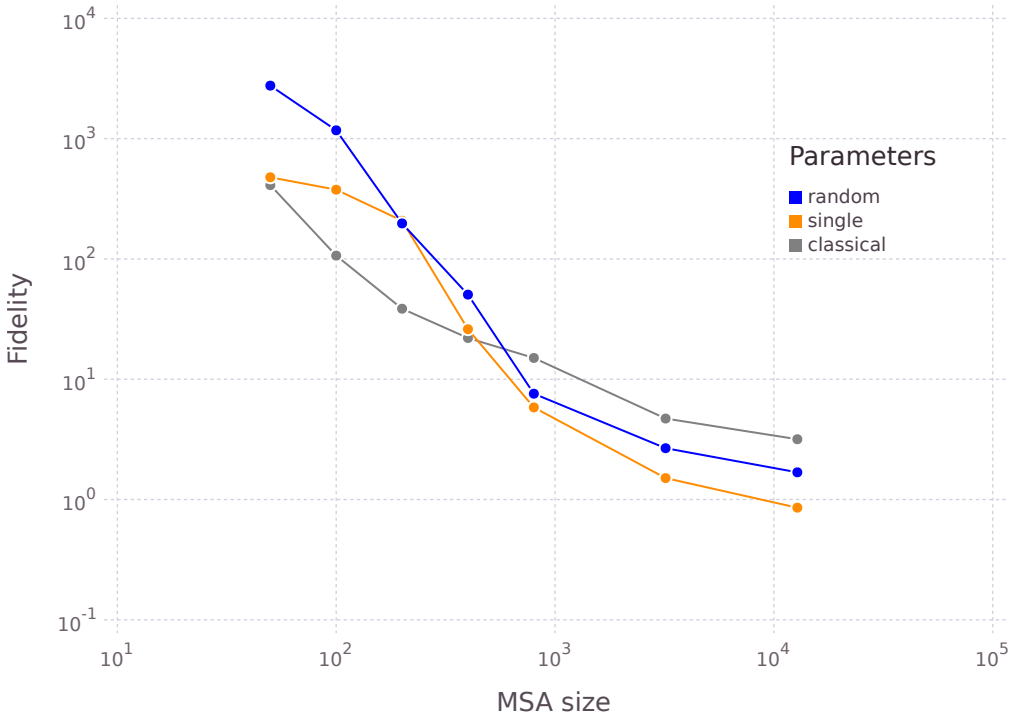

Figure 9: The prediction accuracy increases strongly for increasing sample sizes. Our method outperforms the existing tool for MSAs larger than $3 \times Q^L$. Plotted are the mean values of the fidelity measure introduced in Equation (4.4) over each, a thousand samples generated from a single randomly chosen parameter set (orange), from different random parameter sets (blue), and additionally for results of the classical DCA algorithm for random parameter sets (grey).

5 Conclusion

The estimated underlying probability distribution of amino acid sequences is the basis of our understanding of proteins regarding contact prediction. The PDs are shaped by the Hamiltonian $H(\boldsymbol{\sigma})$ consisting of single site fields and pairwise couplings that govern the physical system. This set of parameters $\{\hat{\boldsymbol{h}}, \hat{\boldsymbol{J}}\}$ is to be fitted according to the sequence data of an MSA. We list the classical Boltzmann distribution (B), the generalization using Tsallis entropy (T), and Rényi entropy (R)

$$p(\boldsymbol{\sigma})_{\mathrm{B}} = \frac{1}{Z_{\mathrm{B}}} \exp\left\{-H(\boldsymbol{\sigma})\right\}, \tag{5.1}$$

$$p(\boldsymbol{\sigma})_{\mathrm{T}} = \frac{1}{Z_{\mathrm{T}}} \left(1 - (1-q)\left\{H(\boldsymbol{\sigma}) - \langle H \rangle\right\}\right)^{\frac{1}{1-q}}, \tag{5.2}$$

$$p(\boldsymbol{\sigma})_{\mathrm{R}} = \frac{1}{Z_{\mathrm{R}}} \left(1 - \frac{q-1}{q}\left\{H(\boldsymbol{\sigma}) - \langle H \rangle\right\}\right)^{\frac{1}{q-1}}. \tag{5.3}$$

The normalization factor Z varies for every distribution and has to be interpreted appropriately (See chapter 2 for exact definitions). Both generalized distributions recover the classical distribution upon $q \to 1$.

From the derived distributions the conditional probabilities to find a certain amino acid at a certain residue, given the rest of the sequence is known, can be calculated. This leads to the pseudo likelihood l_{pseudo}, the product over all residues and all sequences of this conditional probability. l_{pseudo} works as a measure to find the best model describing the input data, resulting in an optimization problem that we solved. With the generalization of entropy, we created a parameter q that qualitatively changes the underlying protein model. This additional degree of freedom is suitable to describe sequencing biases arisen from small sample

sizes. We think that for future tests on real proteins, q will also carry information about the selection pressure the protein is inhibiting, which effectively alters its distributions of sequences away from equilibrium.
From the statistical data generated for a $Q^L = 4^4$ toy model in equilibrium, i.e. $q = 1$, we obtained as main results that the variance of q around 1 largely increases for smaller sample sizes and, on the other hand, the performance of our method exceeds the original DCA algorithm for large sample sizes, while not depending on heuristic distribution reshaping tools, like pseudo counts and reweighting.
Although qualitatively the Tsallis distribution for varying q could be easily mistaken for a Boltzmann-Gibbs distribution with varying temperature, this work proofs a significant difference between the two. While for classical DCA the temperature is effectively considered as a factor in the calculated fields and couplings, the complex power law behavior due to a entropy generalization is a true extension of the theory and gives way to a deeper understanding of proteins and biological processes.

5.1 Prospects

The working TDCA algorithm at hand enables many further studies. In the case of equilibrium, we proved that the q value of the distribution the sequences are drawn from are only well met for large sample sizes. It would be interesting to repeat this test for artificial MSAs drawn from other distributions.
Alternatively, a Gaussian noise could be applied to the underlying fields and couplings when drawing sequences in order to simulate environmental changes to the system. We expect the quality of the added noise to be represented by q.
The main task, however, will be to test real biological data. A promising candidate for first investigations is HIV protease, which is under constant scientific observation [Rhe+03]. This protein inhibits a strong selection pressure due to the comprehensive drug treatment it is subjected to. Therefore we expect results with $q > 1$.
On the other hand, proteins have to be identified, occurring in organisms that have not undergone strong evolutionary changes for billions of years. Such sequences are expected to be drawn from distributions with $q < 1$.
During our research new publications on the field of direct coupling analysis have been released, testing the method also on RNA [DL+15] and as a tool to describe protein-protein interaction [Fei+16], both of which are potential applications for TDCA.

Acknowledgements

The author would like to thank Prof. Kay Hamacher and Michael Schmidt for the idea to this study and the many fruitful discussions throughout the past year. Further thanks to Karin Ibe and Sabrina Müller for critical reading of the manuscript and thoughtful comments.

Appendices

A Additional Results

q for a single parameter set

Sample size	$\varnothing q$	δq	$q_{\min}$	$q_{\max}$
50	0.950	0.022	0.85	1.01
100	0.915	0.034	0.81	1.04
200	0.907	0.073	0.79	1.73
400	0.976	0.109	0.79	1.50
800	0.992	0.034	0.82	1.21
3200	0.999	0.028	0.92	1.10
12800	1.0	0.014	0.95	1.04

q for random parameter sets

Sample size	$\varnothing q$	δq	$q_{\min}$	$q_{\max}$
50	0.950	0.026	0.81	1.01
100	0.906	0.041	0.77	1.02
200	0.918	0.160	0.73	1.93
400	0.973	0.143	0.68	1.83
800	0.989	0.079	0.77	1.26
3200	0.998	0.031	0.89	1.09
12800	1.0	0.016	0.95	1.06

Table 1: Numerical results of a thousand calculations for each sample size with the values for the mean and standard deviation of q and its minimum and maximum value. The upper table inhibits the results for samples generated from a single fixed parameter set and the lower table depicts the results where for every sample a new parameter set was diced. These results are also shown graphically in Figure 7.

	single		random		classical	
Sample size	f_h	f_J	f_h	f_J	f_h	f_J
50	200.605	276.808	855.313	1903.510	70.965	339.881
100	156.259	220.250	436.866	736.631	22.500	84.071
200	87.528	120.372	66.643	130.772	9.545	29.026
400	11.310	14.759	10.847	39.668	6.039	15.978
800	2.318	3.514	2.539	5.043	4.217	10.812
3200	0.467	1.041	0.686	1.985	1.124	3.589
12800	0.223	0.633	0.344	1.339	0.519	2.653

Table 2: Numerical results of a thousand calculations for each sample size with the fidelity of the fields f_h and couplings f_J for a single fixed parameter set and for random parameter sets, each for TDCA and classical DCA. These results are also shown graphically in Figure 9.

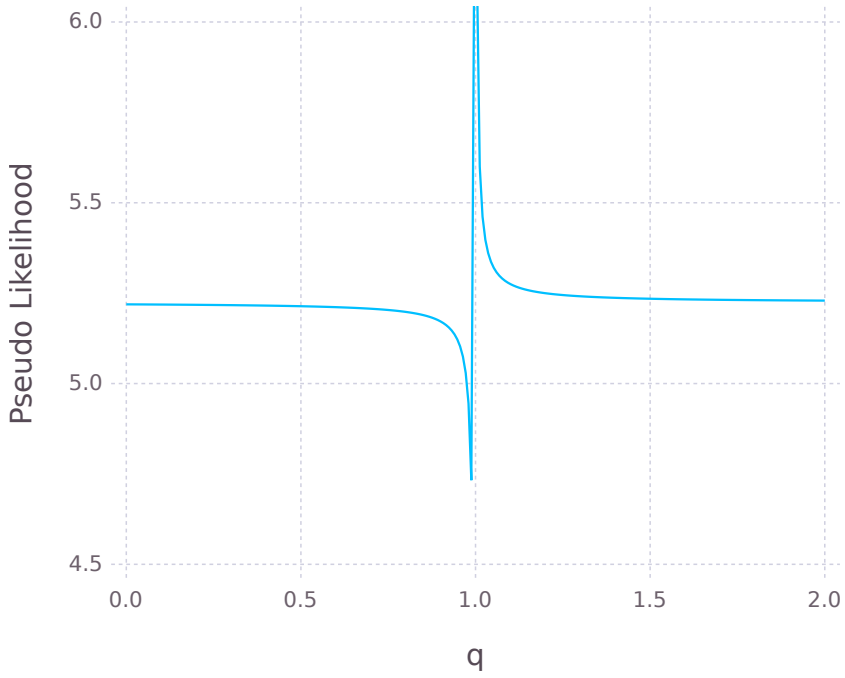

Figure 10: Analysis of the protein Interleukin8 (CXCL8) taken from the PsiCov dataset [Jon+12]. When optimized with the classical DCA algorithm the variation in q for the fixed fields and couplings shows no value more favorable than $q = 1$. The TDCA algorithm also favors $q = 1$, both for step sizes of $\Delta q = 0.1$ and $\Delta q = 0.01$. Possibly even smaller steps are necessary.

In order to supplement the data visualization in Figures 7 and 9, we list the numerical values in Tables 1 and 2, respectively. Additionally, the mentioned analysis of the protein Interleukin8 is visualized in Figure 10. There is no significant difference between l_{pseudo}, whether or not reweighting is used in the calculation. However, the values of the fields and couplings differ slightly.

B Detailed Calculations

In Section 3.1 we introduced expressions for various derivatives of thermodynamic potentials and their linearizations. In this Appendix we show the corresponding calculations in detail. Additionally, the linearization of the exponential constraints from Section 2.3.1 is shown.

B.1 Linear Expansion of the Exponential Expected Value

Starting from the definition of the exponential expected value in Equation (2.16) we find

$$\langle \delta_{\sigma_i k} \rangle_e = \frac{1}{1-q} \ln \left(\sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma}) e^{(1-q)\delta_{\sigma_i k}} \right) \tag{B.1}$$

$$= \frac{1}{1-q} \ln \left(e^{1-q} \sum_{\boldsymbol{\sigma}|\sigma_i=k} p(\boldsymbol{\sigma}) + \sum_{\boldsymbol{\sigma}|\sigma_i \neq k} p(\boldsymbol{\sigma}) \right) \tag{B.2}$$

$$= \frac{1}{1-q} \ln \left(e^{1-q} \sum_{\boldsymbol{\sigma}|\sigma_i=k} p(\boldsymbol{\sigma}) + \left(1 - \sum_{\boldsymbol{\sigma}|\sigma_i=k} p(\boldsymbol{\sigma}) \right) \right) \tag{B.3}$$

$$= \frac{1}{1-q} \ln \left(1 - (1 - e^{1-q}) \sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma}) \delta_{\sigma_i k} \right). \tag{B.4}$$

Let us remind that in all cases $\sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma}) \delta_{\sigma_i k} \leq 1$, thus for q being in the vicinity of 1, even only in the same magnitude as 1, we can Taylor expand the logarithm in (B.4) to linear order around 1, so that

$$\langle \delta_{\sigma_i k} \rangle_e \approx \frac{e^{1-q} - 1}{1-q} \sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma}) \delta_{\sigma_i k} \tag{B.5}$$

$$\approx \sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma}) \delta_{\sigma_i k} \tag{B.6}$$

is still a good approximation.

B.2 Derivatives of the Thermodynamic Potentials

F is the Helmholtz potential, defined by $F = -\ln Z$, where Z is the partition function. Its derivative with respect to a local field, governing the spin k at residue i reads

$$\frac{\partial F(\alpha)}{\partial h_i(k)} = -\frac{1}{Z(\alpha)} \frac{\partial Z(\alpha)}{\partial h_i(k)} \tag{B.7}$$

$$= \frac{\gamma}{Z(\alpha)} \sum_{\boldsymbol{\sigma}} (\delta_{\sigma_i k} - f_i(k)) \mathcal{F}'(H_\alpha^c(\boldsymbol{\sigma})) \mathcal{F}(H_\alpha^c(\boldsymbol{\sigma}))^{\gamma-1} \tag{B.8}$$

$$= \gamma \sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma}) \frac{\mathcal{F}'(H_\alpha^c(\boldsymbol{\sigma}))}{\mathcal{F}(H_\alpha^c(\boldsymbol{\sigma}))} (\delta_{\sigma_i k} - f_i(k)) \tag{B.9}$$

$$= \gamma \sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma}) \frac{\delta_{\sigma_i k} - f_i(k)}{\zeta + H_\alpha^c(\boldsymbol{\sigma})} \tag{B.10}$$

$$\equiv \pi_i(k). \tag{B.11}$$

In this and the following calculations a dash as in $\mathcal{F}'(H_\alpha^c(\boldsymbol{\sigma}))$ is meant in the mathematical sense as differentiation with respect to the argument; as opposed to differentiation with

respect to the field, as it may suggest. With this result we continue with the second partial derivative with respect to a second local field governing spin l at residue j.

$$\frac{\partial^2 F}{\partial h_i(k)\partial h_j(l)} = -\frac{\partial}{\partial h_j(l)}\frac{\gamma}{Z(\alpha)}\sum_{\boldsymbol{\sigma}}(\delta_{\sigma_i k} - f_i(k))\mathcal{F}'(H_\alpha^c(\boldsymbol{\sigma}))\mathcal{F}(H_\alpha^c(\boldsymbol{\sigma}))^{\gamma-1} \tag{B.12}$$

$$= \frac{\gamma(\gamma-1)}{Z(\alpha)}\sum_{\boldsymbol{\sigma}}\{(\delta_{\sigma_i k} - f_i(k))(\delta_{\sigma_j l} - f_j(l))\times$$

$$\times \mathcal{F}'(H_\alpha^c(\boldsymbol{\sigma}))^2\mathcal{F}(H_\alpha^c(\boldsymbol{\sigma}))^{\gamma-2}\} - \pi_i(k)\pi_j(l) \tag{B.13}$$

$$= \gamma(\gamma-1)\sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma})\frac{(\delta_{\sigma_i k} - f_i(k))(\delta_{\sigma_j l} - f_j(l))}{(\zeta + H_\alpha^c(\boldsymbol{\sigma}))^2} - \pi_i(k)\pi_j(l) \tag{B.14}$$

$$\equiv \pi_{ij}(k,l) - \pi_i(k)\pi_j(l), \tag{B.15}$$

In a first step the product rule applies generating to independent single partial derivatives and a combined one. G is the Gibbs potential, defined by a Legendre transformation to the variables $p_i(k)$: $G = -\ln Z + \sum_{k,i} h_i(k)p_i(k)$. We calculate its derivative with respect to α, which is the expansion parameter of the interaction term in a mean field Hamiltonian;

$$\frac{\mathrm{d}G(\alpha)}{\mathrm{d}\alpha} = -\frac{1}{Z(\alpha)}\frac{\partial Z(\alpha)}{\partial \alpha} \tag{B.16}$$

$$= -\frac{1}{Z(\alpha)}\gamma\sum_{\boldsymbol{\sigma}}(H_l(\boldsymbol{\sigma}) - \langle H_l\rangle)\mathcal{F}'(H(\boldsymbol{\sigma},\alpha))\mathcal{F}(H(\boldsymbol{\sigma},\alpha))^{\gamma-1} \tag{B.17}$$

$$= -\gamma\sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma})\frac{H_l(\boldsymbol{\sigma}) - \langle H_l\rangle}{\zeta + H_\alpha^c(\boldsymbol{\sigma})}. \tag{B.18}$$

B.3 Linearization of π_i and π_{ij}

The complicated terms above do not lead to results feasible to deduce a mean field theory, which is why we further linearize for small centralized energies $H_\alpha^c(\boldsymbol{\sigma})$, which are per definition distributed around zero.

$$\pi_i(k) \approx \frac{\gamma}{\zeta}\sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma})(\delta_{\sigma_i k} - f_i(k))(1 - \frac{H_\alpha^c(\boldsymbol{\sigma})}{\zeta}) \tag{B.19}$$

$$= \frac{\gamma}{\zeta}\sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma})\{\delta_{\sigma_i k} - \delta_{\sigma_i k}\frac{H_\alpha^c(\boldsymbol{\sigma})}{\zeta} - f_i(k) + f_i(k)\frac{H_\alpha^c(\boldsymbol{\sigma})}{\zeta}\} \tag{B.20}$$

$$= \frac{\gamma}{\zeta}\{\langle\delta_{\sigma_i k}\rangle - \frac{1}{\zeta}\langle\delta_{\sigma_i k}H_\alpha^c\rangle - f_i(k) + f_i(k)\frac{1}{\zeta}\langle H_\alpha^c\rangle\} \tag{B.21}$$

$$= -\frac{\gamma}{\zeta^2}\langle\delta_{\sigma_i k}H_\alpha^c\rangle \tag{B.22}$$

$$= \frac{\gamma}{\zeta^2}(\langle\delta_{\sigma_i k}\langle H_\alpha\rangle - \delta_{\sigma_i k}H_\alpha\rangle) \tag{B.23}$$

$$= \frac{\gamma}{\zeta^2}\left\{\langle\delta_{\sigma_i k}\rangle\langle H_\alpha\rangle - \langle\delta_{\sigma_i k}H_\alpha\rangle\right\} \tag{B.24}$$

Note that per definition $\langle \delta_{\sigma_i k} \rangle = \sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma}) \delta_{\sigma_i k} = f_i(k)$ and we used the linearity of the expected value. We continue with the second partial derivative of the Helmholtz potential.

$$\pi_{ij}(k,l) \approx \frac{\gamma(\gamma-1)}{\zeta} \sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma})(\delta_{\sigma_i k} - f_i(k))(\delta_{\sigma_j l} - f_j(l))(1 - \frac{2H_\alpha^c(\boldsymbol{\sigma})}{\zeta}) \tag{B.25}$$

$$= \frac{\gamma(\gamma-1)}{\zeta} \{ \langle \delta_{\sigma_i k} \delta_{\sigma_j l} \rangle - \frac{2}{\zeta} \langle \delta_{\sigma_i k} \delta_{\sigma_j l} H_\alpha^c \rangle + f_i(k) f_j(l) - \frac{2}{\zeta} f_i(k) f_j(l) \langle H_\alpha^c \rangle$$
$$- \langle \delta_{\sigma_i k} \rangle f_j(l) - f_i(k) \langle \delta_{\sigma_j l} \rangle + \frac{2}{\zeta} f_j(l) \langle \delta_{\sigma_i k} H_\alpha^c \rangle + \frac{2}{\zeta} f_i(k) \langle \delta_{\sigma_j l} H_\alpha^c \rangle \} \tag{B.26}$$

$$= \frac{\gamma(\gamma-1)}{\zeta} \{ f_{ij}(k,l) - f_i(k) f_j(l)$$
$$- \frac{2}{\zeta} (\langle \delta_{\sigma_i k} \delta_{\sigma_j l} H_\alpha^c \rangle - f_i(k) \langle \delta_{\sigma_j l} H_\alpha^c \rangle - f_j(l) \langle \delta_{\sigma_i k} H_\alpha^c \rangle) \} \tag{B.27}$$

$$= \frac{\gamma(\gamma-1)}{\zeta} \{ f_{ij}(k,l) - f_i(k) f_j(l)$$
$$- \frac{2}{\zeta} (\langle \delta_{\sigma_i k} \delta_{\sigma_j l} H_\alpha \rangle - f_i(k) \langle \delta_{\sigma_j l} H_\alpha \rangle - f_j(l) \langle \delta_{\sigma_i k} H_\alpha \rangle)$$
$$+ \frac{2}{\zeta} (\langle \delta_{\sigma_i k} \delta_{\sigma_j l} \rangle \langle H_\alpha \rangle - f_i(k) \langle \delta_{\sigma_j l} \rangle \langle H_\alpha \rangle - f_j(l) \langle \delta_{\sigma_i k} \rangle \langle H_\alpha \rangle) \} \tag{B.28}$$

$$= \frac{\gamma(\gamma-1)}{\zeta} \{ f_{ij}(k,l) \left(1 + \frac{2}{\zeta} \langle H_\alpha \rangle \right) - f_i(k) f_j(l) \left(1 + \frac{4}{\zeta} \langle H_\alpha \rangle \right)$$
$$- \frac{2}{\zeta} \{ \langle \delta_{\sigma_i k} \delta_{\sigma_j l} H_\alpha \rangle + \langle \delta_{\sigma_i k} H_\alpha \rangle f_j(l) + \langle \delta_{\sigma_j l} H_\alpha \rangle f_i(k) \} \} \tag{B.29}$$

At last the α derivative of the Gibbs potential is linearized in $H_\alpha^c(\boldsymbol{\sigma})$

$$\frac{\mathrm{d}G(\alpha)}{\mathrm{d}\alpha} \approx -\frac{\gamma}{\zeta} \sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma})(H_l(\boldsymbol{\sigma}) - \langle H_l \rangle)(1 - \frac{H_\alpha^c(\boldsymbol{\sigma})}{\zeta}) \tag{B.30}$$

$$= -\frac{\gamma}{\zeta} \{ \langle H_l \rangle - \frac{1}{\zeta} \langle H_l H^c \rangle - \langle H_l \rangle + \frac{1}{\zeta} \langle H_l \rangle \langle H^c \rangle \} \tag{B.31}$$

$$= \frac{\gamma}{\zeta} \{ \langle H_l H_S \rangle - \langle H_l \rangle \langle H_S \rangle + \alpha \left(\langle H_l^2 \rangle - \langle H_l \rangle^2 \right) \}. \tag{B.32}$$

B.4 α Derivative of G

In order to find an expression for the second partial derivative of the Gibbs potential we first take a look at the constituents of its linear Taylor term.

$$\langle H_S H_l \rangle = \sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma}) H_S(\boldsymbol{\sigma}) H_l(\boldsymbol{\sigma}) \tag{B.33}$$

$$= \sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma}) \sum_i h_i(\sigma_i) \sum_{i<j} J_{ij}(\sigma_i, \sigma_j) \tag{B.34}$$

$$= \sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma}) \sum_{i,j<k} h_i(\sigma_i) J_{jk}(\sigma_j, \sigma_k) \tag{B.35}$$

$$= \sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma}) \sum_{i,j<k} \sum_{l,m,n} h_i(l) J_{jk}(m,n) \delta_{\sigma_i l} \delta_{\sigma_j m} \delta_{\sigma_k n} \tag{B.36}$$

$$= \sum_{i,j<k} \sum_{l,m,n} h_i(l) J_{jk}(m,n)\ p_{ijk}(l,m,n) \tag{B.37}$$

$$\overset{\alpha=0}{=} \sum_{i,j<k} \sum_{l,m,n} h_i(l) J_{jk}(m,n)\ p_i(l) p_j(m) p_k(n) \tag{B.38}$$

In the last step we used that for $\alpha = 0$ the multi point distributions factorize to single point distributions.

$$\langle H_S \rangle \langle H_I \rangle = \left(\sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma}) H_S(\boldsymbol{\sigma}) \right) \left(\sum_{\boldsymbol{\sigma}'} p(\boldsymbol{\sigma}') H_I(\boldsymbol{\sigma}') \right) \tag{B.39}$$

$$= \left(\sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma}) \sum_i h_i(\sigma_i) \right) \left(\sum_{\boldsymbol{\sigma}'} p(\boldsymbol{\sigma}') \sum_{i<j} J_{ij}(\sigma'_i, \sigma'_j) \right) \tag{B.40}$$

$$= \left(\sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma}) \sum_{i,l} h_i(l) \delta_{\sigma_i l} \right) \left(\sum_{\boldsymbol{\sigma}'} p(\boldsymbol{\sigma}') \sum_{j<k} \sum_{m,n} J_{jk}(m,n) \delta_{\sigma'_j m} \delta_{\sigma'_k n} \right) \tag{B.41}$$

$$= \left(\sum_{i,l} h_i(l) p_i(l) \right) \left(\sum_{j<k} \sum_{m,n} J_{jk}(m,n)\ p_{jk}(m,n) \right) \tag{B.42}$$

$$= \sum_{i,j<k} \sum_{l,m,n} h_i(l) J_{jk}(m,n)\ p_i(l) p_{jk}(m,n) \tag{B.43}$$

$$\overset{\alpha=0}{=} \sum_{i,j<k} \sum_{l,m,n} h_i(l) J_{jk}(m,n)\ p_i(l) p_j(m) p_k(n) \tag{B.44}$$

Thus we have already shown that the linear term of G in α and $H^c_\alpha(\boldsymbol{\sigma})$ vanishes. If you are interested, this is the derivative of its constituents:

$$\frac{\partial^2}{\partial p_i(k) \partial p_j(l)} \langle H_S H_I \rangle_{\alpha=0}$$

$$= \frac{\partial^2}{\partial p_i(k) \partial p_j(l)} \sum_{m,n,o} \sum_{u,v,w} p_m(u) p_n(v) p_o(w)\ h_m(u) J_{no}(v,w) \tag{B.45}$$

$$= \frac{\partial}{\partial p_i(k)} \sum_{m,n} \sum_{u,v} p_m(u) p_n(v)\ \{h_j(l) J_{mn}(u,v) + h_m(u) J_{nj}(v,l) + h_m(u) J_{jn}(l,v)\} \tag{B.46}$$

$$= \sum_{m,u} p_m(u)\ \{h_m(u) J_{ij}(k,l) + 2 h_i(k) J_{jm}(l,u)\}. \tag{B.47}$$

References

[Sha48] Claude Elwood Shannon. "A mathematical theory of communication". In: *Bell System Technical Journal* 27.3 (1948), pp. 379–423.

[Pot52] Renfrey Burnard Potts. "Some generalized order-disorder transformations". In: *Mathematical proceedings of the cambridge philosophical society*. Vol. 48. 01. Cambridge Univ Press. 1952, pp. 106–109.

[Met+53] Nicholas Metropolis et al. "Equation of state calculations by fast computing machines". In: *The journal of chemical physics* 21.6 (1953), pp. 1087–1092.

[Jay57a] Edwin T. Jaynes. "Information theory and statistical mechanics". In: *Physical review* 106.4 (1957), p. 620.

[Jay57b] Edwin T. Jaynes. "Information theory and statistical mechanics. II". In: *Physical review* 108.2 (1957), p. 171.

[Rén66] Alfréd Rényi. *Wahrscheinlichkeitsrechnung: mit einem Anhang über Informationstheorie*. Vol. 54. Deutscher Verlag der Wissenschaften, 1966.

[Noc80] Jorge Nocedal. "Updating quasi-Newton matrices with limited storage". In: *Mathematics of computation* 35.151 (1980), pp. 773–782.

[Ple82] T. Plefka. "Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model". In: *Journal of Physics A: Mathematical and general* 15.6 (1982), p. 1971.

[Tsa88] Constantino Tsallis. "Possible generalization of Boltzmann-Gibbs statistics". In: *Journal of statistical physics* 52.1-2 (1988), pp. 479–487.

[LN89] Dong C. Liu and Jorge Nocedal. "On the limited memory BFGS method for large scale optimization". In: *Mathematical programming* 45.1-3 (1989), pp. 503–528.

[GY91] Antoine Georges and Jonathan S. Yedidia. "How to expand around mean-field theory using high-temperature expansions". In: *Journal of Physics A: Mathematical and General* 24.9 (1991), p. 2173.

[Tho+97] Julie D. Thompson et al. "The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools". In: *Nucleic acids research* 25.24 (1997), pp. 4876–4882.

[TMP98] Constantino Tsallis, Renio S. Mendes, and Anel R. Plastino. "The role of constraints within generalized nonextensive statistics". In: *Physica A: Statistical Mechanics and its Applications* 261.3 (1998), pp. 534–554.

[DS+99] R. P. Di Sisto et al. "General thermostatistical formalisms, invariance under uniform spectrum translations, and Tsallis q-additivity". In: *Physica A: Statistical Mechanics and its Applications* 265.3 (1999), pp. 590–613.

[LY99] Elliott H. Lieb and Jakob Yngvason. "The physics and mathematics of the second law of thermodynamics". In: *Physics Reports* 310.1 (1999), pp. 1–96.

[Mar+00] S. Martınez et al. "Tsallis' entropy maximization procedure revisited". In: *Physica A: Statistical Mechanics and its Applications* 286.3 (2000), pp. 489–502.

[Abe+01] Sumiyoshi Abe et al. "Nonextensive thermodynamic relations". In: *Physics Letters A* 281.2 (2001), pp. 126–130.

[Abe02] Sumiyoshi Abe. "Stability of Tsallis entropy and instabilities of Rényi and normalized Tsallis entropies: A basis for q-exponential distributions". In: *Physical Review E* 66.4 (2002), p. 046134.

[Yam02] Takuya Yamano. "Some properties of q-logarithm and q-exponential functions in Tsallis statistics". In: *Physica A: Statistical Mechanics and its Applications* 305.3–4 (2002), pp. 486 –496. issn: 0378-4371.

[Rhe+03] Soo-Yon Rhee et al. "Human immunodeficiency virus reverse transcriptase and protease sequence database". In: *Nucleic acids research* 31.1 (2003), pp. 298–303.

[Bas04] A. G. Bashkirov. "On maximum entropy principle, superstatistics, power-law distribution and Renyi parameter". In: *Physica A: Statistical Mechanics and its Applications* 340.1 (2004), pp. 153–162.

[FMP05a] G. L. Ferri, S. Martinez, and A. Plastino. "Equivalence of the four versions of Tsallis's statistics". In: *Journal of Statistical Mechanics: Theory and Experiment* 2005.04 (2005), P04009.

[FMP05b] GL Ferri, S Martínez, and A Plastino. "The role of constraints in Tsallis' nonextensive treatment revisited". In: *Physica A: Statistical Mechanics and its Applications* 347 (2005), pp. 205–220.

[KS05] Sergei Kucherenko and Yury Sytsko. "Application of deterministic low-discrepancy sequences in global optimization". In: *Computational Optimization and Applications* 30.3 (2005), pp. 297–318.

[Mas05] Marco Masi. "A step beyond Tsallis and Rényi entropies". In: *Physics Letters A* 338.3 (2005), pp. 217–224.

[WS05] T. Wada and A. M. Scarfone. "A non self-referential expression of Tsallis' probability distribution function". In: *The European Physical Journal B-Condensed Matter and Complex Systems* 47.4 (2005), pp. 557–561.

[Com06] Wikimedia Common. *IL8 Solution Structure.rsh.png*. `https://upload.wikimedia.org/wikipedia/commons/f/f6/IL8_Solution_Structure.rsh.png`. 2006.

[Git] *GitHub*. `https://github.com/`. 2008.

[CSN09] Aaron Clauset, Cosma Rohilla Shalizi, and Mark E. J. Newman. "Power-law distributions in empirical data". In: *SIAM review* 51.4 (2009), pp. 661–703.

[MM09] Marc Mezard and Thierry Mora. "Constraint satisfaction problems and neural networks: A statistical physics perspective". In: *Journal of Physiology-Paris* 103.1 (2009), pp. 107–113.

[RAH09] Yasser Roudi, Erik Aurell, and John A. Hertz. "Statistical physics of pairwise probability models". In: *Frontiers in computational neuroscience* 3 (2009).

[SM09] Vitor Sessak and Rémi Monasson. "Small-correlation expansions for the inverse Ising problem". In: *Journal of Physics A: Mathematical and Theoretical* 42.5 (2009), p. 055001.

[Wei+09] Martin Weigt et al. "Identification of direct residue contacts in protein–protein interaction by message passing". In: *Proceedings of the National Academy of Sciences* 106.1 (2009), pp. 67–72.

[ZRM09] Royce K. P. Zia, Edward F. Redish, and Susan R. McKay. "Making sense of the Legendre transform". In: *American Journal of Physics* 77.7 (2009), pp. 614–622.

[RWL+10] Pradeep Ravikumar, Martin J. Wainwright, John D. Lafferty, et al. "High-dimensional Ising model selection using L1-regularized logistic regression". In: *The Annals of Statistics* 38.3 (2010), pp. 1287–1319.

[CM11] S. Cocco and R. Monasson. "Adaptive Cluster Expansion for Inferring Boltzmann Machines with Noisy Data". In: *Phys. Rev. Lett.* 106 (9 2011), p. 090601.

[Mor+11] Faruck Morcos et al. "Direct-coupling analysis of residue coevolution captures native contacts across many protein families". In: *Proceedings of the National Academy of Sciences* 108.49 (2011), E1293–E1301.

[AE12] Erik Aurell and Magnus Ekeberg. "Inverse Ising Inference Using All the Data". In: *Phys. Rev. Lett.* 108 (9 2012), p. 090201.

[GVL12] Gene H. Golub and Charles F. Van Loan. *Matrix computations*. Vol. 3. JHU Press, 2012.

[Jon+12] David T. Jones et al. "PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments". In: *Bioinformatics* 28.2 (2012), pp. 184–190. eprint: `http://bioinformatics.oxfordjournals.org/content/28/2/184.full.pdf+html`.

[NB12a] H. Chau Nguyen and Johannes Berg. "Bethe–Peierls approximation and the inverse Ising problem". In: *Journal of Statistical Mechanics: Theory and Experiment* 2012.03 (2012), P03004.

[NB12b] H. Chau Nguyen and Johannes Berg. "Mean-Field Theory for the Inverse Ising Problem at Low Temperatures". In: *Phys. Rev. Lett.* 109 (5 2012), p. 050602.

[RT12] Federico Ricci-Tersenghi. "The Bethe approximation for solving the inverse Ising problem: a comparison with other inference methods". In: *Journal of Statistical Mechanics: Theory and Experiment* 2012.08 (2012), P08015.

[Jul] *The Julia Language*. `http://julialang.org/`. 2012.

[Eke+13] Magnus Ekeberg et al. "Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models". In: *Physical Review E* 87.1 (2013), p. 012707.

[Fin+13] Robert D. Finn et al. "Pfam: the protein families database". In: *Nucleic acids research* (2013), gkt1223.

[Bar+14] J. P. Barton et al. "Large pseudocounts and L2-norm penalties are necessary for the mean-field inference of Ising and Potts models". In: *Physical Review E* 90.1 (2014), p. 012132.

[CB14] Michele Castellana and William Bialek. "Inverse Spin Glass and Related Maximum Entropy Problems". In: *Phys. Rev. Lett.* 113 (11 2014), p. 117204.

[DRT14] Aurélien Decelle and Federico Ricci-Tersenghi. "Pseudolikelihood decimation algorithm improving the inference of the interaction network in a general class of ising models". In: *Physical review letters* 112.7 (2014), p. 070603.

[EHA14] Magnus Ekeberg, Tuomo Hartonen, and Erik Aurell. "Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences". In: *Journal of Computational Physics* 276 (2014), pp. 341–356.

[DL+15] Eleonora De Leonardis et al. "Direct-Coupling Analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction". In: *Nucleic Acids Research* 43.21 (2015), pp. 10444–10455. eprint: `http://nar.oxfordjournals.org/content/43/21/10444.full.pdf+html`.

[SMS15] Richard R. Stein, Debora S. Marks, and Chris Sander. "Inferring Pairwise Interactions from Biological Data Using Maximum-Entropy Probability Models". In: *PLoS computational biology* 11.7 (2015).

[Fei+16] Christoph Feinauer et al. “Inter-Protein Sequence Co-Evolution Predicts Known Physical Interactions in Bacterial Ribosomes and the Trp Operon”. In: *PLoS ONE* 11.2 (Feb. 2016), pp. 1–18.