IOP PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND THEORETICAL

J. Phys. A: Math. Theor. **41** (2008) 285004 (50pp) doi:10.1088/1751-8113/41/28/285004

A solvable model of the genesis of amino-acid sequences via coupled dynamics of folding and slow-genetic variation

S Rabello[1], A C C Coolen[2,3], C J Pérez-Vicente[4] and F Fraternali[3]

[1] Department of Mathematics, Imperial College London, South Kensington Campus, London SW7 2AZ, UK
[2] Department of Mathematics, King's College London, The Strand, London WC2R 2LS, UK
[3] Randall Division of Cell and Molecular Biophysics, King's College London, New Hunt's House, London SE1 1UL, UK
[4] Departament de Fisica Fonamental, Facultat de Fisica, Universitat de Barcelona, 08028 Barcelona, Spain

E-mail: s.rabello@imperial.ac.uk, ton.coolen@kcl.ac.uk, conrad@ffn.ub.es and franca.fraternali@kcl.ac.uk

Received 22 February 2008, in final form 20 May 2008
Published 19 June 2008
Online at stacks.iop.org/JPhysA/41/285004

Abstract

We study the coupled dynamics of primary and secondary structures formation (i.e. slow-genetic sequence selection and fast folding) in the context of a solvable microscopic model that includes both short-range steric forces and long-range polarity-driven forces. Our solution is based on the diagonalization of replicated transfer matrices, and leads in the thermodynamic limit to explicit predictions regarding phase transitions and phase diagrams at genetic equilibrium. The predicted phenomenology allows for natural physical interpretations, and finds satisfactory support in numerical simulations.

PACS numbers: 61.41.+e, 75.10.Nr

1. Introduction

The constituent monomers of protein-type hetero-polymers, the amino acids of which there exist about 20 in nature, are composed of a common backbone and a differentiating side chain, and are bound via a peptide bond. These units are connected sequentially to form a polypeptide chain. The sequence of connected amino acids defines the so-called primary structure of the chain. Given the primary structure, the mechanical degrees of freedom of the polypeptide chain are rotation angles at the junctions of adjacent amino acids. They allow proteins to fold into relatively simple repetitive local arrangements (the 'secondary structures', such as α-helices or β-sheets) which then combine into more complicated global

arrangements in 3D (the 'tertiary structure'). The folding process is controlled by various combinations of forces, such as those induced by mutual interactions between the amino-acid side chains (steric forces, Van der Waals forces), by interactions between side chains and the polymer's backbone (hydrogen and sulfur bonds) and by interactions between the amino-acid side chains and the surrounding solvent (polarity-induced forces and hydrogen bonds). For comprehensive reviews on the physics of the interactions governing the folding of proteins, see e.g. [1, 2]. Apart from 'chaperone' effects (the influence of specialized proteins), it was discovered [3] that the dynamics of the folding process is for most proteins determined solely by their primary structure. Since polypeptide chains can vary in length from a few tens to tens of thousands of monomers, there is an enormous number of possible sequences. Yet only a tiny fraction of these (the actual biologically functional proteins) will represent chains that fold into a unique reproducible tertiary structure, or three-dimensional 'conformation', which determines its biological function.

The protein folding problem is how to predict this conformation (the native state) of a protein, given its primary structure. It remains one of the most challenging unsolved problems in biology. Its solution would have a big impact on medicine. The physicist's strategy in this field (as opposed to bio-informatics approaches based on simulation, see e.g. [4] for a recent review) is to try to understand the main physical mechanisms that drive the one-to-one correspondence between amino-acid sequence and the native state. Normally, this is attempted via simple quantitative mathematical models that capture the essential phenomenology of folding and lend themselves to statistical mechanical analysis [5–7] and/or are easily simulated numerically [8–11]. In the language of thermodynamics and statistical mechanics, it is believed that if a protein spontaneously reaches its native state at physiological conditions of temperature and pressure, its free-energy landscape must possess a unique stable minimum [12]. However, calculating free-energy landscapes for biologically functional proteins is non-trivial, because of the frustration induced by the local steric constraints in combination with the effective interactions via polarity and hydrogen bonds, especially in view of the heterogeneity of the amino-acid sequences. In addition we would like to understand the folding pathway that ensures a protein's fast approach to its native state in physiological conditions, by avoiding kinetic traps and minimizing the various potential frustration effects [13, 14]. Random amino-acid sequences do not fold into unique conformations, i.e. they have more complicated multi-valley free-energy landscapes, so one concludes that those sequences that correspond to proteins have been selected genetically on the basis of their associated free-energy landscapes [15, 16].

There is little consensus yet as to what is the main driving force in the folding process. Some believe the hydrophobic–hydrophilic effect (i.e. hydrophobic side chains try to avoid contact with the solvent, while hydrophilic side chains seek to be in contact with it) to be the dominant factor in secondary and tertiary structure formation [15–18], with steric constraints enforcing further microscopic specificity, and hydrogen bonds providing a locking mechanism [19]. Others believe the folding to be mainly driven by the formation of intra-molecular (or peptidic) hydrogen bonds on top of hydrogen bonding between side chains and the solvent [20]. Most physicists' studies either resort to models similar to self-avoiding walks on regular lattices [21, 22] (usually via graph counting and numerical simulations), or focus on generic properties of (free) energy landscapes [23–25], or try to exploit the one-dimensional nature of the polypeptide chains [26–28]. In either case, in virtually all studies the amino-acid sequences are regarded as frozen disorder, over which appropriate averages are calculated (in statics of the free energy per monomer, in dynamics of the moment-generating dynamical functional). This implies that the sequences at hand must be 'typical' within an appropriate ensemble of sequences, which presents us with a serious fundamental problem. Amino-acid

sequences of proteins are far from random: they have been carefully selected during evolution on the basis of their functionality and their ability to lead to reproducible folds. Thus, one either has to define an ensemble of amino-acid sequences on the basis of the known primary sequences of real proteins that are being collected in biological databases, which removes the possibility to carry out disorder averages in the mathematical theory analytically, or one has to find a way to capture the essence of the observed biological sequences (as opposed to random ones) in simple mathematical formulae. Although some analytical studies did involve non-random sequences, the sequence statistics were usually not connected to folding quality as such [29, 30].

There is an alternative strategy in the statistical mechanical modeling of interacting many-particle systems with non-random disorder, which was followed successfully in the past, for e.g. neural networks (where the synaptic connections between neurons represent the disorder) [31–34] and for a simple mean-field hetero-polymer model [35]. Rather than averaging over all amino-acid sequences (subject perhaps to experimentally determined constraints), one combines the process of secondary structure generation (folding) with a slow-evolutionary process for the amino-acid sequences (which represents the genetic selection of free-energy landscapes) and one couples these two processes in a biologically acceptable way. One can then try to solve for the 'slow' process upon assuming adiabatic separation of the two time scales, using the so-called finite-n replica theory. This results in solvable models describing structure generation in polypeptide chains with amino-acid sequences that are no longer random, but selected in a manner that correlates with the folding process, without having been required to capture the sequence statistics in a formula. It is encouraging that we know from previous studies such as [31–35] that in such models the impact of the slow-genetic process is indeed generally to drive the systems away from multi-valley energy landscapes toward single-valley ones.

In the present paper we take the next step in this research programme, whereas [35] involved a simplified model with only polarity-induced mean-field forces, here we develop a theory for the coupled dynamics of (fast) folding and (slow) sequence selection on the basis of the more precise Hamiltonian introduced in [26], which also includes short-range steric forces along the chain. At a technical level our problem requires the diagonalization of replicated transfer matrices, for which efficient methods have been developed only recently [36–39]. We apply these diagonalization methods to the present model, within the ergodic (i.e. replica symmetric, RS) ansatz, and show how they lead in the thermodynamic limit to closed equations for non-trivial order parameters. In the context of protein folding one expects the RS ansatz to be appropriate. In finite-dimensional replica calculations replica symmetry is known to break down only for small values of the replica dimension n, i.e. at high genetic noise levels, whereas here our interest is mostly in the regime of low-genetic noise levels. Second, given the robustness and reproducibility of proteins' secondary and tertiary structures one must assume these systems to operate in an ergodic regime. Third, at a mathematical level, our present order parameter equations will involve only quantities with a single replica index, giving yet another indication that RS should hold. After first recovering the solutions of the order parameter equations in various known limits, we focus on the biologically most realistic regime of sequence selection at zero genetic noise levels, namely $n \to \infty$, where we extract the non-trivial phase phenomenology and derive phase diagrams analytically. We find many interesting phase transitions, both continuous and discontinuous, and remanence effects, all of which can be understood and explained on physical grounds. This is followed by a numerical analysis of the order parameter equations for nonzero genetic noise levels, and by tests of the theoretical predictions against numerical simulations of the coupled sequence selection and folding processes. Within the limitations imposed by finite size and finite relaxation time

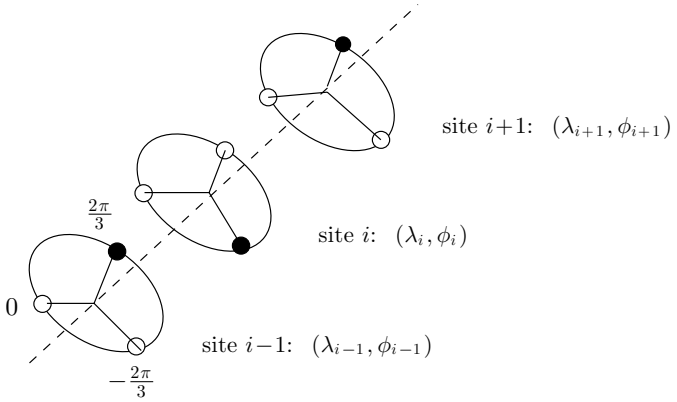

Figure 1. Illustration of the chemical and mechanical degrees of freedom in our model. At each site i of the chain we have a discrete variable λ_i which specifies the local amino acid *type*, and a residue angle ϕ_i which defines its physical *location* relative to the one-dimensional polymer chain axis (the 'backbone', drawn as a dashed line). In this example the number of possible orientations of each residue is three. The black blobs represent locations occupied by residues. The primary structure of the polymer (its chemical composition) is thus defined by $(\lambda_1, \ldots, \lambda_N)$, and the secondary structure by $(\phi_1, \ldots, \phi_N)$. Both types of variables are assumed to evolve in time, although on widely separated time scales.

effects, we find a satisfactory agreement between our theoretical predictions and the numerical simulations.

2. Model definitions

2.1. The folding and sequence selection processes

Our model inherits much of its initial features from [26], and represents the amino-acid cores as nodes in a one-dimensional chain. The global conformal state of the system is defined by N successive angles $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_N) \in \Omega^N$ of amino-acid residues, relative to the chain's backbone. Here $\Omega = \{0, 2\pi/q, 4\pi/q, \ldots, (q-1)2\pi/q\} \subset [0, 2\pi)$, where $q \in \mathbb{N}$. The simplified picture is that of residues being able to rotate (with constraints, and limited to q discrete positions) in a plane perpendicular to the chain's axis. The primary structure (the amino-acid sequence) is written as $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N)$, with $\lambda_i \in \{1, \ldots, \Lambda\}$ denoting the residue species at position i in the chain (with $\Lambda = 20$ for real proteins). See also figure 1. In contrast to [26], however, the primary sequence will here not be drawn at random, but will be generated by an appropriate genetic selection process; this improves the biological realism of the model, but will change and complicate the mathematics significantly. We will therefore only include monomer-solvent polarity forces and steric forces, leaving out hydrogen bonds for now. Furthermore, we refine the Hamiltonian used in [26] to take into account the effect of the polymer's overall polarity balance on its ability to exhibit predominantly hydrophilic surface residues and hydrophobic core residues; for models with fixed primary sequences as in [26] this would add an irrelevant constant to the energy, but for models such as the present where the monomer sequences evolve in time this energy contribution will exert sequence selection pressure with significant consequences. In many of our calculations we will also choose $q = 2$, i.e. limit the residue angles to $\phi_i \in \{0, \pi\}$. This prevents us from having to

generalize the diagonalization methods of [36, 37], which would probably require a separate study in itself. Thus, for a given realization of the primary sequence $\boldsymbol{\lambda}$, the folding process is assumed to be governed by the following Hamiltonian:

$$H_{\mathrm{f}}(\phi|\boldsymbol{\lambda}) = -\frac{J_p}{N}\sum_{ij}\xi(\lambda_i)\xi(\lambda_j)\delta_{\phi_i,\phi_j} - J_s\sum_i \cos[(\phi_{i+1}-\phi_i)-(\phi_i-\phi_{i-1})-a(\lambda_i)]. \tag{1}$$

$\xi(\lambda)\in\mathbb{R}$ measures the polarity of residue λ (with $\xi>0$ indicating hydrophobicity and $\xi_i<0$ indicating hydrophilicity). The first term in (1) favors conformations where hydrophobic and hydrophilic avoid identical orientations, since this makes it easier for the polymer to find a fold that shields its hydrophobic residues from the solvent while exposing its hydrophilic ones. The second term represents in a simplified manner the effects of steric forces, characterizing each residue λ by a winding 'distortion' angle $a(\lambda)$ for successive residue rotations. If $a(\lambda_i)=0$, then residue i will prefer to have an angle ϕ_i such that torsion along the chain is homogeneous, i.e. $\phi_{i+1}-\phi_i=\phi_i-\phi_{i-1}$. The energies $J_p>J_s>0$ control the relative impact of each contribution. For a fixed sequence one can define the partition function $Z_{\mathrm{f}}(\boldsymbol{\lambda})$ and the free energy $F_{\mathrm{f}}(\boldsymbol{\lambda})$ for the equilibrium state of the folding process at temperature $T_{\mathrm{f}}=\beta^{-1}$ (in units where the Boltzmann constant equals $k_B=1$):

$$\mathcal{Z}_{\mathrm{f}}(\boldsymbol{\lambda}) = \sum_{\phi}\exp[-\beta H_{\mathrm{f}}(\phi|\boldsymbol{\lambda})], \tag{2}$$

$$F_{\mathrm{f}}(\boldsymbol{\lambda}) = -\beta^{-1}\log\mathcal{Z}_{\mathrm{f}}(\boldsymbol{\lambda}). \tag{3}$$

It will be convenient to characterize the relevant chemical characteristics of amino acids by the distribution

$$w(\xi,\eta) = \frac{1}{\Lambda}\sum_{\lambda=1}^{\Lambda}\delta[\xi-\xi(\lambda)]\delta[\eta-\cos[a(\lambda)]]. \tag{4}$$

In principle, (4) reduces to a collection of 20 points in the (ξ,η) plane, but since it is impossible to extract their precise locations directly from available data (see section 2.2) one has to rely on sensible proxies[5]. As there is no obvious structural physical/chemical link between residue polarity and geometric (steric) properties, we assume statistical independence, i.e. $w(\xi,\eta)=w(\xi)w(\eta)$ (this will also induce welcome simplifications later). Typical simple choices for $w(\xi)$ would be $w(\xi)=\epsilon\delta(\xi)+\frac{1}{2}(1-\epsilon)[\delta(\xi-1)+\delta(\xi+1)]$ or $w(\xi)=\frac{1}{2}\theta[1-\xi]\theta[1+\xi]$. Note that we may always choose the maximum polarity to be one, since alternative values can be absorbed into the definition of the parameter J_p. For $w(\eta)$, natural choices would be $w(\eta)=\pi^{-1}\int_0^{\pi}\mathrm{d}a\,\delta[\eta-\cos(a)]=\pi^{-1}[1-\arccos^2(\eta)]^{-1/2}\theta[1-\eta]\theta[1+\eta]$ or $w(\eta)=\frac{1}{2}\theta[1-\eta]\theta[1+\eta]$. Here the allowed value range $[-1,1]$ is enforced by the physical meaning of η.

We now follow [35] and complement the folding process by an adiabatically slow-stochastic evolutionary selection process for the amino-acid sequences. The assumption is that this selection results from an interplay between the demands that (i) a sequence must lead to a unique and easily reproducible equilibrium conformation for its associated folding process, and (ii) the resulting structure is useful to the organism (e.g. it can act as a catalyst

[5] We will in fact find that in the limit $N\to\infty$ the only dependence of the system's phase diagram on the amino-acid characteristics $\{\xi(\lambda),a(\lambda)\}$ is via the distribution $w(\xi,\eta)$, and involves only qualitative properties of this distribution, such as symmetries and its finite support.

of some metabolic or proteomic cellular reaction). If one takes the further step to quantify the quality of an equilibrium conformation by the value of the folding free energy $F_{\rm f}(\boldsymbol{\lambda})$ (i.e. taking 'low-free energy' as a proxy for 'more reproducible'), together with the direct energetic cost $V(\boldsymbol{\lambda})$ of not having strictly hydrophilic 'surface residues' and strictly hydrophobic 'core residues', and if one assumes that biological usefulness can be measured by some utility potential $U(\boldsymbol{\lambda})$, then the evolutionary process can be viewed as the stochastic minimization of an effective Hamiltonian for amino-acid sequences that takes the form

$$H_{\rm eff}(\boldsymbol{\lambda}) = U(\boldsymbol{\lambda}) + V(\boldsymbol{\lambda}) - \beta^{-1} \log \mathcal{Z}_{\rm f}(\boldsymbol{\lambda}). \tag{5}$$

If the stochastic minimization is of the Glauber or Monte Carlo type, the evolutionary process will evolve itself to a Boltzmann-type equilibrium state, namely $P_\infty(\boldsymbol{\lambda}) \propto \exp[-\tilde{\beta} H_{\rm eff}(\boldsymbol{\lambda})]$, where $\tilde{\beta}$ measures the (inverse) noise level in the genetic selection[6]. Our combined model (fast-folding and slow-genetic sequence selection) is thus solved in equilibrium by calculating the associated effective free energy per monomer

$$\begin{aligned} f_N &= -\frac{1}{\tilde{\beta} N} \log \sum_{\boldsymbol{\lambda}} \mathrm{e}^{-\tilde{\beta} H_{\rm eff}(\boldsymbol{\lambda})} \\ &= -\frac{1}{n\beta N} \log \sum_{\boldsymbol{\lambda}} [\mathcal{Z}_{\rm f}(\boldsymbol{\lambda})]^n \, \mathrm{e}^{-n\beta[U(\boldsymbol{\lambda})+V(\boldsymbol{\lambda})]} \end{aligned} \tag{6}$$

with the noise level ratio $n = \tilde{\beta}/\beta$. As in [31–35], this expression can be evaluated via the replica formalism, where n is first taken to be integer and the result is subsequently continued to non-integer values. Note that in this type of model the replica dimension has a clear physical meaning as the ratio of temperatures. For $n \to 0$ we recover the free energy of a system with quenched random amino-acid sequences; for $n = 1$ we have that of an annealed model, whereas for $n \to \infty$ the sequence selection becomes strictly deterministic. In contrast to previous coupled dynamics studies, however, here we have not only mean-field forces but also short-range ones: the steric interactions in (1). The replica calculation will therefore be quite different.

In this paper we limit ourselves for mathematical convenience to sequence functionality potentials of the simple form $U(\boldsymbol{\lambda}) = \sum_i u(\lambda_i)$. Similarly we choose the energetic penalty $V(\boldsymbol{\lambda})$ on hydrophobic surface residues or hydrophilic core residues to be a function only of the polarity balance $k(\boldsymbol{\lambda}) = N^{-1} \sum_i \xi(\lambda_i)$, putting $V(\boldsymbol{\lambda}) = J_g N v(k(\boldsymbol{\lambda}) - k^\star)$ with a function $v(k)$ that is minimal for $k = 0$, where $k^\star$ represents the 'optimal' polarity balance that would give a protein with strictly hydrophilic surface residues and strictly hydrophilic core residues (which one expects to be close to zero). This form for $V(\boldsymbol{\lambda})$ would emerge naturally if all amino acids were to have similar values of $|\xi(\lambda_i)|$. The implicit assumption is that if a polarity balance $k(\boldsymbol{\lambda})$ is energetically favorable, i.e. close to $k^\star$, then the protein will be able to find a fold that realizes the desired geometric separation of core versus surface residues. We will discuss the mathematical consequences of making alternative choices in section 8. Since for $N \to \infty$ chain boundary effects must vanish, we also choose periodic boundary conditions and take N even (for mathematical reasons which will become clear later).

[6] Another way to see why $P_\infty(\boldsymbol{\lambda}) \propto \exp[-\tilde{\beta} H_{\rm eff}(\boldsymbol{\lambda})]$ is a natural evolutionary equilibrium state is to image having real-valued $\boldsymbol{\lambda}$, evolving according to a Langevin equation in which the deterministic force is minus the gradient of the energy $H_{\rm f}(\boldsymbol{\lambda}) + U(\boldsymbol{\lambda}) + V(\boldsymbol{\lambda})$. Given adiabatic separation of folding and evolution time scales, one can then integrate out the fast variables (the conformation angles) and find the Boltzmann state for the sequences $\boldsymbol{\lambda}$ with effective Hamiltonian (5). See e.g. [35] for details.

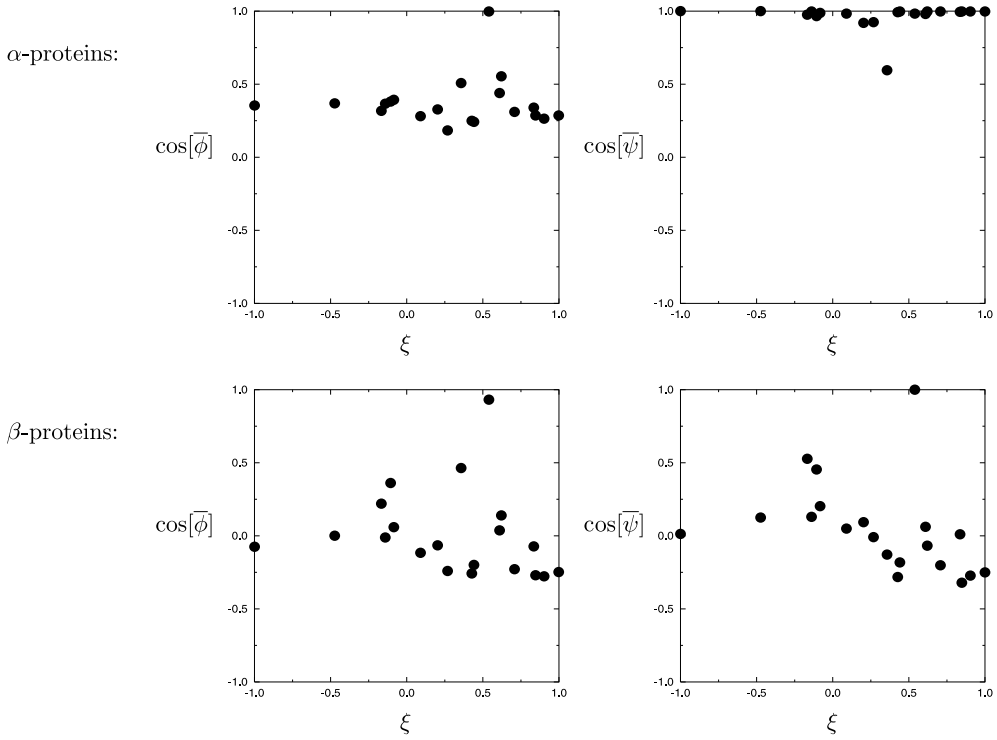

Figure 2. Diagrams showing each of the 20 amino acids as a point in the plane, with the horizontal coordinate giving its polarity value (taken from [43], and normalized to the range $[-1, 1]$), and with the horizontal coordinate giving the cosine of an average conformation angle (averaged over all proteins of a given class). Left: averages calculated for conformation angle ϕ; right: averages calculated for conformation angle ψ. Top row: averaging over all α-proteins for which structures are available; bottom row: averaging over all β-proteins for which structures are available. All conformation data were extracted from [41, 42].

2.2. Relation between model assumptions and biological reality

Here we discuss some of the assumptions and definitions of our model in the light of experimental evidence from real proteins. Our choice for a single-angle representation of the mechanical degrees of freedom of a monomer was motivated by our desire to limit the mathematical complexity, although our methods would also apply if we were to work with the conventional two conformation angles (ϕ, ψ). In fact, there is evidence [40] to suggest that the conventional two-angle representation is redundant, and that only one newly defined torsion angle is needed per amino acid to specify a protein's conformation. If we insist on identifying the single-site degrees of freedom in our model with one of the standard conformation angles (ϕ, ψ), we have to choose the one that matches our statistical assumptions best. To do this, we have calculated for individual amino acids the average of the observed conformation angles (ϕ, ψ) over all occurrences of this amino acid in the database of known protein structures (the SCOP database [41, 42]), which resulted in the graphs of figure 2, where we plot the cosines of the average conformation angles of all 20 amino acids together with their polarity values (according to the Eisenberg scale, taken from [43], and normalized linearly to the range $[-1, 1]$). Both conformation angles (ϕ, ψ) give averages that have cosines of both signs, both are biased toward positive values; however, the bias is more extreme in the case of ψ. Since there is no such bias in our theory, the most suitable conformation angle to

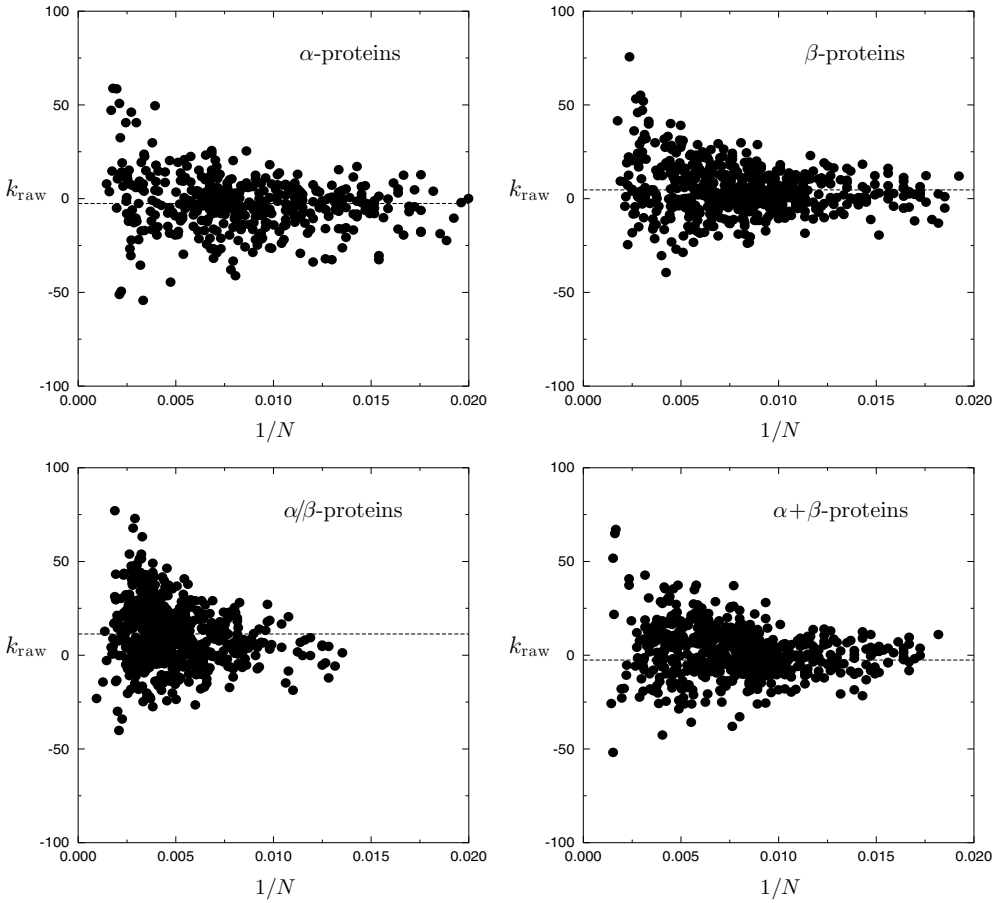

Figure 3. Diagrams showing all proteins in database [42], organized into the four main protein families, as points in the plane, with the horizontal coordinate giving their inverse size N^{-1} and with the vertical coordinate giving their *average* polarity value $k_{\rm raw} = N^{-1}\sum_i \xi_{i,\rm raw}$ along the chain (where the polarities $\xi_{i,\rm raw}$ of the constituent residues are taken directly from [43], without normalization to $[-1, 1]$). Dashed horizontal lines indicate the average overall polarity level found within each protein class.

correspond to the orientation degrees of freedom in our model appears to be ϕ. In the same figure we can also see that there is no obvious correlation between polarity characteristics and steric characteristics. In our model this is assumed to be a property of the amino acids, and we will find in our analysis that neither the primary structure generation nor the secondary structure generation introduces any such correlations. Finally, let us turn to the postulated preferred average polarity of any amino-acid chain (which was used in our phenomenological Hamiltonian), purely on the basis of the energetic need to shield hydrophobic residues from the solvent and to expose hydrophilic ones. There is certainly evidence for the link between the average polarity of a sequence and the surface-exposure pattern of the associated protein structure [44]. If we plot all those proteins for which primary structure data are available as points in a plane, with the inverse size $1/N$ as a horizontal coordinate and the average polarity as a vertical coordinate, we obtain figure 3. This figure supports strongly the existence of an energetically preferred average polarity $k^\star$, with a value close to zero in rescaled polarity units $\xi \in [-1, 1]$.

3. Replica analysis of the model

For integer n one can write the nth power of the folding partition function $\mathcal{Z}_{\rm f}(\boldsymbol{\lambda})$ in (6) in terms of n replicas of the original system, to be labeled by $\alpha = 1, \ldots, n$. If the sum over the sequences $\boldsymbol{\lambda}$ is carried out before the sum over conformations, one finds an effective theory in which the n replicas are coupled:

$$f_N = -\frac{1}{n\beta N} \log \sum_{\phi^1,\ldots,\phi^n} \mathrm{e}^{-\beta \mathcal{H}(\phi^1,\ldots,\phi^n)} \tag{7}$$

$$\mathcal{H}(\cdots) = -\frac{1}{\beta} \log \sum_{\boldsymbol{\lambda}} \mathrm{e}^{-\beta \sum_\alpha H_{\rm f}(\phi^\alpha|\boldsymbol{\lambda}) - n\beta[U(\boldsymbol{\lambda})+V(\boldsymbol{\lambda})]}. \tag{8}$$

For $\beta \to 0$ (infinite temperature) we have $\beta\mathcal{H}(\cdots) \to -N \log \Lambda$ and the free energy retains only entropic terms, namely $\lim_{\beta\to 0}(\beta f_N) = -\log q - n^{-1} \log \Lambda$. Upon using (1), and inserting $\sum_\phi \delta_{\phi,\phi_i^\alpha}$ into the polarity term of the folding energy, we can work out the effective Hamiltonian (8). If we introduce appropriate integrals over δ-functions (written in integral representation) to isolate the quantities $N^{-1}\sum_i \xi(\lambda_i)\delta_{\phi,\phi_i^\alpha}$, namely

$$1 = \int \frac{\mathrm{d}z_{\alpha\phi}\, \mathrm{d}\hat{z}_{\alpha\phi}}{2\pi} \mathrm{e}^{\mathrm{i}\hat{z}_{\alpha\phi}[z_{\alpha\phi} - N^{-1}\sum_i \xi(\lambda_i)\delta_{\phi,\phi_i^\alpha}]} \tag{9}$$

where $\phi^\alpha = (\phi_1^\alpha, \ldots, \phi_N^\alpha)$, we can carry out the sum over $\boldsymbol{\lambda}$ in (8) and find, with the abbreviation $\mathbf{z} = \{z_{\alpha\phi}\}$,

$$\begin{aligned} -\frac{\beta}{N}\mathcal{H}(\cdots) &= \frac{1}{N} \log \sum_{\boldsymbol{\lambda}} \mathrm{e}^{-n\beta[\sum_i u(\lambda_i) + N J_g v(k(\boldsymbol{\lambda})-k^\star)] + \frac{\beta J_p}{N}\sum_\phi \sum_\alpha [\sum_i \xi(\lambda_i)\delta_{\phi,\phi_i^\alpha}]^2} \\ &\quad \times \mathrm{e}^{\beta J_s \sum_{i\alpha} \cos[\phi_{i+1}^\alpha + \phi_{i-1}^\alpha - 2\phi_i^\alpha - a(\lambda_i)]} \\ &= \frac{1}{N} \log \int \frac{\mathrm{d}\mathbf{z}\, \mathrm{d}\hat{\mathbf{z}}}{(2\pi/\beta N)^{qn}} \mathrm{e}^{\beta N[\mathrm{i}\sum_{\alpha\phi}\hat{z}_{\alpha\phi}z_{\alpha\phi} + J_p\mathbf{z}^2 - nJ_g v(\frac{1}{n}\sum_{\alpha\phi} z_{\alpha\phi} - k^\star)]} \\ &\quad \times \prod_i \left\{ \sum_\lambda \mathrm{e}^{-n\beta u(\lambda) - \mathrm{i}\beta\xi(\lambda)\sum_{\alpha\phi}\hat{z}_{\alpha\phi}\delta_{\phi,\phi_i^\alpha} + \beta J_s \sum_\alpha \cos[\phi_{i+1}^\alpha + \phi_{i-1}^\alpha - 2\phi_i^\alpha - a(\lambda)]} \right\}. \end{aligned} \tag{10}$$

Inserting this into (7) leads to an expression for the asymptotic free energy per monomer $f = \lim_{N\to\infty} f_N$ that can be evaluated by the steepest descent. Upon eliminating the conjugate integration variables $\{\hat{z}_{\alpha\phi}\}$ by variation of $\{z_{\alpha\phi}\}$, giving $\mathrm{i}\hat{z}_{\alpha\phi} = J_g v'\big(\frac{1}{n}\sum_{\alpha\phi} z_{\alpha\phi} - k^\star\big) - 2J_p z_{\alpha\phi}$, and upon defining the replicated single-site vectors $\boldsymbol{\phi}_i = \big(\phi_i^1, \ldots, \phi_i^n\big)$ the result takes the form $f = \mathrm{extr}_{\mathbf{z}}\varphi_n(\mathbf{z})$ with

$$\begin{aligned} n\varphi_n(\mathbf{z}) &= J_p \sum_{\alpha\phi} z_{\alpha\phi}^2 + nJ_g \left[v\left(\frac{1}{n}\sum_{\alpha\phi} z_{\alpha\phi} - k^\star\right) - \left(\frac{1}{n}\sum_{\alpha\phi} z_{\alpha\phi}\right) v'\left(\frac{1}{n}\sum_{\alpha\phi} z_{\alpha\phi} - k^\star\right)\right] \\ &\quad - \frac{1}{\beta}\log\Lambda - \lim_{N\to\infty}\frac{1}{\beta N}\log \sum_{\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_N} \prod_i M[\boldsymbol{\phi}_{i-1}, \boldsymbol{\phi}_i, \boldsymbol{\phi}_{i+1}|\mathbf{z}], \end{aligned} \tag{11}$$

$$\begin{aligned} M[\boldsymbol{\phi}_{i-1}, \boldsymbol{\phi}_i, \boldsymbol{\phi}_{i+1}|\mathbf{z}] &= \frac{1}{\Lambda}\sum_{\lambda=1}^{\Lambda} \mathrm{e}^{\beta\xi(\lambda)\sum_\alpha [2J_p z_{\alpha\phi_i^\alpha} - J_g v'(\frac{1}{n}\sum_{\alpha\phi} z_{\alpha\phi} - k^\star)]} \\ &\quad \times \mathrm{e}^{\beta J_s \sum_\alpha \cos[\phi_{i+1}^\alpha + \phi_{i-1}^\alpha - 2\phi_i^\alpha - a(\lambda)] - n\beta u(\lambda)}. \end{aligned} \tag{12}$$

We recognize in (11) and (12) a replicated transfer matrix product embedded within a mean-field calculation, and conclude that this model is therefore in principle solvable. The only amino-acid characteristics that affect the folding process are its polarity $\xi(\lambda)$ and steric angle $a(\lambda)$, so we will from now on choose the single site functionality potential to have the form $u(\lambda) = \mu\xi(\lambda) + \nu\cos[a(\lambda)]$ (where μ and ν are control parameters).

3.1. The case $q = 2$

Our calculations become significantly simpler and more transparent for $q = 2$. Here, after a uniform basis rotation, the allowed residue angles are $\phi_i \in \{-\pi/2, \pi/2\}$, which can be written in terms of Ising spin variables $\sigma_i \in \{-1, 1\}$ as $\phi_i = \sigma_i\pi/2$. We transform the $2n$ remaining replicated order parameters, which can be written as $z_{\alpha\pm}$, into new order parameters $m_\alpha = z_{\alpha+} - z_{\alpha-}$ and $k_\alpha = z_{\alpha+} + z_{\alpha-}$. Our equations will now involve the replicated spin variables $\boldsymbol{\sigma}_i = (\sigma_i^1, \ldots, \sigma_i^n)$, and the cosine term in the exponent of the transfer matrix simplifies to $\sigma_{i+1}^\alpha\sigma_{i-1}^\alpha\cos[a(\lambda)]$. With the short-hands $\mathbf{m} = (m_1, \ldots, m_n)$ and $\mathbf{k} = (k_1, \ldots, k_n)$, $\langle g(\xi)\rangle_\xi = \int \mathrm{d}\xi\, w(\xi)g(\xi)$, and $\langle g(\eta)\rangle_\eta = \int \mathrm{d}\eta\, w(\eta)g(\eta)$, our previous expressions (11) and (12) take the form

$$n\varphi_n(\mathbf{m},\mathbf{k}) = \frac{1}{2}J_p(\mathbf{k}^2+\mathbf{m}^2) + nJ_g\left[v\left(\frac{1}{n}\sum_\alpha k_\alpha - k^\star\right) - \left(\frac{1}{n}\sum_\alpha k_\alpha\right)v'\left(\frac{1}{n}\sum_\alpha k_\alpha - k^\star\right)\right]$$
$$-\frac{1}{\beta}\log\Lambda - \lim_{N\to\infty}\frac{1}{\beta N}\log\sum_{\boldsymbol{\sigma}_1,\ldots,\boldsymbol{\sigma}_N}\prod_i M[\boldsymbol{\sigma}_{i-1},\boldsymbol{\sigma}_i,\boldsymbol{\sigma}_{i+1}|\mathbf{m},\mathbf{k}], \tag{13}$$

$$M[\boldsymbol{\sigma}_{i-1},\boldsymbol{\sigma}_i,\boldsymbol{\sigma}_{i+1}|\mathbf{m},\mathbf{k}] = \left\langle \mathrm{e}^{\beta\xi[J_p\sum_\alpha(k_\alpha+m_\alpha\sigma_i^\alpha)-n\mu-nJ_gv'(\frac{1}{n}\sum_\alpha k_\alpha-k^\star)]}\right\rangle_\xi$$
$$\times\left\langle \mathrm{e}^{\beta\eta[J_s\boldsymbol{\sigma}_{i+1}\cdot\boldsymbol{\sigma}_{i-1}-n\nu]}\right\rangle_\eta. \tag{14}$$

The disconnection inside $M[\cdots|\cdots]$ of the factor involving $\boldsymbol{\sigma}_i$ from that involving $\boldsymbol{\sigma}_{i+1}\cdot\boldsymbol{\sigma}_{i-1}$ allows us to rewrite $\varphi_n(\mathbf{m},\psi)$ into a more convenient form, with a new replicated transfer matrix $\boldsymbol{\Gamma}(\mathbf{m},\mathbf{k})$ that involves only the two sites $i-1$ and $i+1$:

$$\prod_i M[\boldsymbol{\sigma}_{i-1},\boldsymbol{\sigma}_i,\boldsymbol{\sigma}_{i+1}|\mathbf{m},\mathbf{k}] = \prod_i \Gamma_{\boldsymbol{\sigma}_{i-1},\boldsymbol{\sigma}_{i+1}}(\mathbf{m},\mathbf{k}), \tag{15}$$

where

$$\Gamma_{\boldsymbol{\sigma}\boldsymbol{\sigma}'}(\mathbf{m},\mathbf{k}) = \left\langle \mathrm{e}^{\beta\eta[J_s\boldsymbol{\sigma}\cdot\boldsymbol{\sigma}'-n\nu]}\right\rangle_\eta\left\langle \mathrm{e}^{\beta\xi[J_p(\sum_\alpha k_\alpha+\mathbf{m}\cdot\boldsymbol{\sigma})-n\mu-nJ_gv'(\frac{1}{n}\sum_\alpha k_\alpha-k^\star)]}\right\rangle_\xi. \tag{16}$$

Since N is even and we have periodic boundaries, even sites thereby disconnect from odd sites. The trace in φ_n is now in leading order for large N expressed in the usual manner in terms of the largest eigenvalue $\lambda(\mathbf{m},\mathbf{k})$ of the matrix $\boldsymbol{\Gamma}(\mathbf{m},\mathbf{k})$:

$$\lim_{N\to\infty}\frac{1}{N}\log\sum_{\boldsymbol{\sigma}_1,\ldots,\boldsymbol{\sigma}_N}\prod_i M[\cdots|\cdots]$$
$$= \lim_{N\to\infty}\frac{2}{N}\log\mathrm{Tr}[\boldsymbol{\Gamma}^{N/2}(\mathbf{m},\mathbf{k})] = \log\lambda(\mathbf{m},\mathbf{k}). \tag{17}$$

In fact, the specific dependence of $\varphi_n(\mathbf{m},\mathbf{k})$ on $\mathbf{k}$ via (14) is such that all its saddle points will have $\mathbf{k} = k(1,\ldots,1)$. This reduces the number of order parameters from $2n$ to $n+1$. We now have $f = \mathrm{extr}_{\mathbf{m},k}\varphi_n(\mathbf{m},k)$, with

$$\varphi_n(\mathbf{m},\mathbf{k}) = \frac{1}{2}J_p\left(\frac{\mathbf{m}^2}{n}+k^2\right) + J_g[v(k-k^\star) - kv'(k-k^\star)] - \frac{1}{n\beta}\log\Lambda$$
$$-\frac{1}{n\beta}\log\lambda(\mathbf{m},k), \tag{18}$$

$$\Gamma_{\sigma\sigma'}(\mathbf{m}, k) = \left\langle e^{\beta\eta[J_s\boldsymbol{\sigma}\cdot\boldsymbol{\sigma}'-n\nu]}\right\rangle_\eta \left\langle e^{n\beta\xi[J_p(k+n^{-1}\mathbf{m}\cdot\boldsymbol{\sigma})-\mu-J_g v'(k-k^\star)]}\right\rangle_\xi. \tag{19}$$

Our problem has been reduced to the diagonalization of the $2^n \times 2^n$ replicated transfer matrix (19). This matrix can be simplified to a form analyzed in [36–39] upon making the so-called replica symmetric (RS) ansatz, which is equivalent to assuming ergodicity. Since the order parameters in the present model have at most one replica index, one expects RS to be exact at all temperatures. Now one has $m_\alpha = m$ for all α, which simplifies our solution to $f = \mathrm{extr}_{m,k}\varphi_{\mathrm{RS}}(m, k)$ in which

$$\varphi_n(m, k) = \frac{1}{2}J_p(m^2 + k^2) + J_g[v(k - k^\star) - kv'(k - k^\star)] - \frac{\log\Lambda + \log\lambda_{\mathrm{RS}}(m, k)}{\beta n}, \tag{20}$$

where $\lambda_{\mathrm{RS}}(m, k)$ is the largest eigenvalue of

$$\Gamma^{\mathrm{RS}}_{\sigma\sigma'}(m, k) = \left\langle e^{\beta\eta[J_s\boldsymbol{\sigma}\cdot\boldsymbol{\sigma}'-n\nu]}\right\rangle_\eta \left\langle e^{n\beta\xi[J_p(k+\frac{m}{n}\sum_\alpha\sigma_\alpha)-\mu-J_g v'(k-k^\star)]}\right\rangle_\xi. \tag{21}$$

Working out the saddle-point equations for $\{m, k\}$ from (20) leads us to

$$m = \frac{1}{\beta n J_p}\frac{\partial}{\partial m}\log\lambda_{\mathrm{RS}}(m, k), \tag{22}$$

$$k = \frac{1}{\beta n[J_p - J_g v''(k - k^\star)]}\frac{\partial}{\partial k}\log\lambda_{\mathrm{RS}}(m, k). \tag{23}$$

An alternative (but equivalent) form for our order parameter equations that does not require differentiation of $\lambda_{\mathrm{RS}}(m, k)$ is obtained if we extremize $\varphi_n(m, k)$ at the stage where it is still expressed in terms of a trace of powers of the matrix $\Gamma_{\mathrm{RS}}(m, k)$, namely

$$\begin{aligned} m &= \frac{2}{\beta n J_p}\lim_{N\to\infty}\frac{\partial}{\partial m}\log\mathrm{Tr}\left[\Gamma^{N/2}_{\mathrm{RS}}(m, k)\right] \\ &= \frac{1}{\beta n J_p}\lim_{N\to\infty}\frac{\mathrm{Tr}\left[\frac{\partial}{\partial m}\Gamma_{\mathrm{RS}}(m, k)\cdot\Gamma^{N/2}_{\mathrm{RS}}(m, k)\right]}{\mathrm{Tr}\left[\Gamma^{N/2}_{\mathrm{RS}}(m, k)\right]}, \end{aligned} \tag{24}$$

$$\begin{aligned} k &= \frac{2}{\beta n}\frac{\lim_{N\to\infty}\frac{\partial}{\partial k}\log\mathrm{Tr}\left[\Gamma^{N/2}_{\mathrm{RS}}(m, k)\right]}{J_p - J_g v''(k - k^\star)} \\ &= \frac{1}{\beta n}\lim_{N\to\infty}\frac{\mathrm{Tr}\left[\frac{\partial}{\partial k}\Gamma_{\mathrm{RS}}(m, k)\cdot\Gamma^{N/2}_{\mathrm{RS}}(m, k)\right]}{[J_p - J_g v''(k - k^\star)]\mathrm{Tr}\left[\Gamma^{N/2}_{\mathrm{RS}}(m, k)\right]}. \end{aligned} \tag{25}$$

Upon working out the partial derivatives of $\Gamma_{\mathrm{RS}}(m, k)$, and upon writing the left and right eigenvectors of $\Gamma_{\mathrm{RS}}(m, k)$ corresponding to the largest eigenvalue as $\{u^{\mathrm{L}}_\sigma\}$ and $\{u^{\mathrm{R}}_\sigma\}$, the limit $N \to \infty$ can be taken. To avoid unwieldy equations we drop the explicit mentioning of the arguments (m, k) for quantities such as λ_{RS}, $\{u^{\mathrm{L}}_\sigma\}$ or $\{u^{\mathrm{R}}_\sigma\}$ from now on; the formulae should make this dependence clear. Using the replica permutation invariance of RS equations, the result can be written as

$$m = \frac{\sum_{\sigma\sigma'}u^{\mathrm{L}}_\sigma\sigma_1 Y_{\sigma\sigma'}u^{\mathrm{R}}_{\sigma'}}{\lambda_{\mathrm{RS}}\sum_\sigma u^{\mathrm{L}}_\sigma u^{\mathrm{R}}_\sigma}, \tag{26}$$

$$k = \frac{\sum_{\sigma\sigma'}u^{\mathrm{L}}_\sigma Y_{\sigma\sigma'}u^{\mathrm{R}}_{\sigma'}}{\lambda_{\mathrm{RS}}\sum_\sigma u^{\mathrm{L}}_\sigma u^{\mathrm{R}}_\sigma}, \tag{27}$$

where

$$Y_{\sigma\sigma'} = \left\langle e^{\beta\eta[J_s\boldsymbol{\sigma}\cdot\boldsymbol{\sigma}'-n\nu]}\right\rangle_\eta \left\langle \xi\, e^{n\beta\xi[J_p(k+\frac{m}{n}\sum_\alpha\sigma_\alpha)-\mu-J_g v'(k-k^\star)]}\right\rangle_\xi. \tag{28}$$

Finally, the physical meaning of the order parameters m and k, expressed in terms of the original variables $\{\sigma_i, \lambda_i\}$ and averages over the equilibrated coupled relaxation processes, is found to be (see appendix A):

$$m = \lim_{N\to\infty}\frac{1}{N}\sum_i \langle\langle \xi(\lambda_i)\sigma_i\rangle\rangle, \tag{29}$$

$$k = \lim_{N\to\infty}\frac{1}{N}\sum_i \langle\langle \xi(\lambda_i)\rangle\rangle \tag{30}$$

(with double brackets $\langle\langle\cdots\rangle\rangle$ denoting equilibrium averages over both the fast secondary structure formation process and the slow-sequence selection process). Within the present model we may interpret $m = 0$, where the equilibrium amino-acid residue orientations are uncorrelated with amino-acid species, as describing a 'swollen' state where secondary structure fails to develop (although, as we will find, for $m = 0$ there could still be phase transitions in terms of the amino-acid statistics, as measured by the order parameter k). States with $m \neq 0$ would exhibit secondary, and by construction (via the polarity term in the folding Hamiltonian) also tertiary structure, so should be described as 'collapsed' states.

3.2. Solution of the replicated eigenvalue problem

It was argued in [36] that the left and right eigenvectors $\{u^{\mathrm{L}}_{\boldsymbol{\sigma}}\}$ and $\{u^{\mathrm{R}}_{\boldsymbol{\sigma}}\}$ corresponding to the largest eigenvalue of matrices of the class (21) are of the following form:

$$u^{\mathrm{R}}_{\boldsymbol{\sigma}} = \int \mathrm{d}x\, \Phi(x)\, e^{\beta x\sum_\alpha\sigma_\alpha}, \tag{31}$$

$$u^{\mathrm{L}}_{\boldsymbol{\sigma}} = \int \mathrm{d}y\, \Psi(y)\, e^{\beta y\sum_\alpha\sigma_\alpha}. \tag{32}$$

Inserting (31) and (32) into the right/left eigenvalue equations $\sum_{\boldsymbol{\sigma}'}\Gamma^{\mathrm{RS}}_{\boldsymbol{\sigma}\boldsymbol{\sigma}'}u^{\mathrm{R}}_{\boldsymbol{\sigma}'} = \lambda_{\mathrm{RS}}u^{\mathrm{R}}_{\boldsymbol{\sigma}}$ and $\sum_{\boldsymbol{\sigma}'}\Gamma^{\mathrm{RS}}_{\boldsymbol{\sigma}'\boldsymbol{\sigma}}u^{\mathrm{L}}_{\boldsymbol{\sigma}'} = \lambda_{\mathrm{RS}}u^{\mathrm{L}}_{\boldsymbol{\sigma}}$, followed by use of the identity $g(\pm1) = \exp[\beta(B \pm A)]$ with

$$A = \frac{1}{2\beta}\log[g(1)/g(-1)], \qquad B = \frac{1}{2\beta}\log[g(1)g(-1)] \tag{33}$$

leads us to a re-formulation of our eigenvalue problems in terms of integral operators, where the role of n has changed from controlling the dimension of the problem (limited to integer values) to that of a simple parameter that can be continued to the real line:

$$\lambda_{\mathrm{RS}}\Phi(x) = \int \mathrm{d}x'\,\Lambda_\Phi(x,x')\Phi(x'), \tag{34}$$

$$\lambda_{\mathrm{RS}}\Psi(x) = \int \mathrm{d}x'\,\Lambda_\Psi(x,x')\Psi(x'). \tag{35}$$

With help of the short-hands

$$A(x,y) = \frac{1}{\beta}\tanh^{-1}[\tanh(\beta x)\tanh(\beta y)], \tag{36}$$

$$B(x, y) = \frac{1}{2\beta} \log[4 \cosh[\beta(x + y)] \cosh[\beta(x - y)]], \tag{37}$$

one can write the kernels in (34) and (35) (of which again we seek the largest eigenvalue) as

$$\Lambda_\Phi(x, x') = \left\langle\!\left\langle \delta[x - \xi J_p m - A(x', \eta J_s)]\, \mathrm{e}^{n\beta[B(x', \eta J_s) + \xi(J_p k - \mu - J_g v'(k - k^\star)) - \nu\eta]} \right\rangle\!\right\rangle_{\xi,\eta}, \tag{38}$$

$$\Lambda_\Psi(x, x') = \left\langle\!\left\langle \delta[x - A(x' + \xi J_p m, \eta J_s)]\, \mathrm{e}^{n\beta[B(x' + \xi J_p m, \eta J_s) + \xi(J_p k - \mu - J_g v'(k - k^\star)) - \nu\eta]} \right\rangle\!\right\rangle_{\xi,\eta}. \tag{39}$$

Both kernels $\Lambda_\Phi(x, x')$ and $\Lambda_\Psi(x, x')$ take only non-negative values, so the eigenvalue problems (34) and (35) support solutions where $\Phi(x) \geqslant 0$ and $\Psi(x) \geqslant 0$ for all $x \in \mathbb{R}$. We may then normalize these functions according to $\int \mathrm{d}x\, \Phi(x) = \int \mathrm{d}x\, \Psi(x) = 1$ and interpret both, in view of (31) and (32), as field distributions. A consequence of this normalization convention is that we obtain two relatively simple (and equivalent) expressions for the eigenvalue λ_{RS} upon integration of (34) and (35) over x:

$$\lambda_{\mathrm{RS}} = \int \mathrm{d}x\, \Phi(x) \left\langle\!\left\langle \mathrm{e}^{n\beta[B(x, \eta J_s) + \xi(J_p k - \mu - J_g v'(k - k^\star)) - \nu\eta]} \right\rangle\!\right\rangle_{\xi,\eta}, \tag{40}$$

$$\lambda_{\mathrm{RS}} = \int \mathrm{d}x\, \Psi(x) \left\langle\!\left\langle \mathrm{e}^{n\beta[B(x + \xi J_p m, \eta J_s) + \xi(J_p k - \mu - J_g v'(k - k^\star)) - \nu\eta]} \right\rangle\!\right\rangle_{\xi,\eta}. \tag{41}$$

Given the normalized solutions $\Psi(x)$ and $\Psi(x)$ of (34) and (35) with the largest eigenvalue, which will generally have to be obtained by numerical iteration, we can work out the remaining contributions to our order parameter equations (26) and (27), such as

$$\sum_\sigma u_\sigma^{\mathrm{L}} u_\sigma^{\mathrm{R}} = 2^n \int \mathrm{d}x\, \mathrm{d}x'\, \Phi(x') \Psi(x) \cosh^n[\beta(x + x')], \tag{42}$$

$$\begin{aligned} \sum_{\sigma\sigma'} u_\sigma^{\mathrm{L}} Y_{\sigma\sigma'} u_{\sigma'}^{\mathrm{R}} = 2^n \int \mathrm{d}x\, \mathrm{d}x'\, \Phi(x') \Psi(x) \Big\langle\!\Big\langle \xi\, \mathrm{e}^{n\beta[B(x', \eta J_s) + \xi(J_p k - \mu - J_g v'(k - k^\star)) - \eta\nu]} \\ \times \cosh^n[\beta(x + \xi J_p m + A(x', \eta J_s))] \Big\rangle\!\Big\rangle_{\xi,\eta}, \end{aligned} \tag{43}$$

$$\begin{aligned} \sum_{\sigma\sigma'} u_\sigma^{\mathrm{L}} \sigma_1 Y_{\sigma\sigma'} u_{\sigma'}^{\mathrm{R}} = 2^n \int \mathrm{d}x\, \mathrm{d}x'\, \Phi(x') \Psi(x) \Big\langle\!\Big\langle \xi \tanh[\beta(x + \xi J_p m + A(x', \eta J_s))] \\ \times \mathrm{e}^{n\beta[B(x', \eta J_s) + \xi(J_p k - \mu - J_g v'(k - k^\star)) - \eta\nu]} \\ \times \cosh^n[\beta(x + \xi J_p m + A(x', \eta J_s))] \Big\rangle\!\Big\rangle_{\xi,\eta}. \end{aligned} \tag{44}$$

3.3. Simplified form of the theory

Equations (26), (27), (34) and (35) (where we need the eigenfunctions with the largest eigenvalue) together with the supporting expressions (38)–(44) constitute a closed set of equations for the RS order parameters $\{m, k, \Phi(x), \Psi(x)\}$ of our model. We now simplify this set further. First we define the following polarity probability density:

$$p(\xi) = \frac{w(\xi)\, \mathrm{e}^{n\beta\xi(J_p k - \mu - J_g v'(k - k^\star))}}{\int \mathrm{d}\xi' w(\xi')\, \mathrm{e}^{n\beta\xi'(J_p k - \mu - J_g v'(k - k^\star))}}. \tag{45}$$

It represents the amino-acid statistics that would have been observed in the absence of the fast process (see appendix A). If we normalize the eigenfunctions $\{\Phi(x), \Psi(x)\}$ according to

$\int \mathrm{d}x\ \Phi(x) = \int \mathrm{d}x\ \Psi(x) = 1$ we find that they are to be solved from

$$\Phi(x) = \int \mathrm{d}\xi\ p(\xi) \left\{ \frac{\int \mathrm{d}x'\ \Phi(x') \int \mathrm{d}\eta\ w(\eta)\delta[x - \xi J_p m - A(x', \eta J_s)]\ \mathrm{e}^{n\beta[B(x',\eta J_s)-\nu\eta]}}{\int \mathrm{d}x'\ \Phi(x') \int \mathrm{d}\eta\ w(\eta)\ \mathrm{e}^{n\beta[B(x',\eta J_s)-\nu\eta]}} \right\}, \quad (46)$$

$$\Psi(x) = \frac{\int \mathrm{d}x' \int \mathrm{d}\xi\ p(\xi)\Psi(x' - \xi J_p m) \int \mathrm{d}\eta\ w(\eta)\delta[x - A(x', \eta J_s)]\ \mathrm{e}^{n\beta[B(x',\eta J_s)-\nu\eta]}}{\int \mathrm{d}x' \int \mathrm{d}\xi\ p(\xi)\Psi(x' - \xi J_p m) \int \mathrm{d}\eta\ w(\eta)\ \mathrm{e}^{n\beta[B(x',\eta J_s)-\nu\eta]}}. \quad (47)$$

The variables x in $\Phi(x)$ and $\Psi(x)$ have the dimension (in spin language) of fields, so $\Phi(x)$ and $\Psi(x)$ must represent field distributions. In fact they are connected in a very explicit way: they can be expressed in terms of each other via

$$\Psi(x) = \frac{\int \mathrm{d}x'\ \Phi(x') \int \mathrm{d}\eta\ w(\eta)\delta[x - A(x', \eta J_s)]\ \mathrm{e}^{n\beta[B(x',\eta J_s)-\nu\eta]}}{\int \mathrm{d}x'\ \Phi(x') \int \mathrm{d}\eta\ w(\eta)\ \mathrm{e}^{n\beta[B(x',\eta J_s)-\nu\eta]}}, \quad (48)$$

$$\Phi(x) = \int \mathrm{d}\xi\ p(\xi)\Psi(x - J_p m\xi). \quad (49)$$

One proves these statements by substituting (48) into (47) and (49) into (46), which shows in either case that both sides of the respective equations are identical. The remaining eigenvalue problem (47) is still non-trivial, but some properties of its solution(s) can be established easily. First, it follows from $|\tanh(\beta A(x', \eta J_s))| = |\tanh(\beta x')\tanh(\beta\eta J_s)| \leqslant \tanh(\beta|\eta|J_s)$ that any solution $\Psi(x)$ must have $\Psi(x) = 0$ for $|x| > J_s \max_{\eta, w(\eta)>0} |\eta|$. Second, as soon as $J_s > 0$ and $J_p m \neq 0$ there cannot be solutions of the trivial form $\Psi(x) = \delta(x - x^\star)$ for finite n. This is clear upon inserting $\Psi(x') = \delta(x' - x^\star)$ into the right-hand side of (47): for $J_s > 0$ and $J_p m \neq 0$ there will *always* be multiple values of $A(x^\star, \eta J_s)$ (since η and ξ take multiple nonzero values), so it is impossible for the right-hand side of (47) to produce a δ-function.

We can now eliminate the distribution $\Phi(x)$ and its eigenvalue problem from our theory, and reduce our order parameter equations to a set involving $\{m, k, \Psi(x)\}$ only. The function Ψ is still to be solved from the eigenvalue equation (47), whereas our two scalar order parameter equations can now be made to take the transparent form

$$m = \int \mathrm{d}\xi\ \mathrm{d}h W(h, \xi)\xi \tanh[\beta h], \quad (50)$$

$$k = \int \mathrm{d}\xi\ \mathrm{d}h W(h, \xi)\xi, \quad (51)$$

with the joint equilibrium distribution $W(h, \xi)$ of local effective fields and polarities:

$$W(h, \xi) = \frac{p(\xi)\cosh^n[\beta h] \int \mathrm{d}x\ \Psi(x)\Psi(h - x - J_p m\xi)}{\int \mathrm{d}\xi'\ \mathrm{d}h' p(\xi')\cosh^n[\beta h'] \int \mathrm{d}x\ \Psi(x)\Psi(h' - x - J_p m\xi')}. \quad (52)$$

Upon calculating the equilibrium distribution

$$\pi(\xi, \eta) = \lim_{N\to\infty} \frac{1}{N} \sum_i \langle\langle \delta[\xi - \xi(\lambda_i)]\delta[\eta - \cos[a(\lambda_i)]]\rangle\rangle \quad (53)$$

(see appendix A) one finds that $\pi(\xi, \eta) = \pi(\xi)\pi(\eta)$, and that $\pi(\xi) = \int \mathrm{d}h W(h, \xi)$. The equilibrium distributions $\pi(\xi)$ and $\pi(\eta)$ will generally differ from the prior distributions $w(\xi)$ and $w(\eta)$ that would be found upon simply drawing amino acids at each site randomly and independently. However, the factorization $\pi(\xi, \eta) = \pi(\xi)\pi(\eta)$ tells us that, although it impacts on amino-acid statistics, in the present model the sequence selection process does not induce correlations between polarity and steric angles.

Given a solution of equations (47), (50) and (51) we can evaluate whether it is the physical one (i.e. that with the lowest free energy) by calculating (20), which now takes the simple

form

$$\varphi = \frac{1}{2}J_p(m^2+k^2) + J_g[v(k-k^\star) - kv'(k-k^\star)] - \frac{\log\Lambda}{\beta n}$$
$$- \frac{1}{\beta n}\log\int \mathrm{d}\xi\; w(\xi)\,\mathrm{e}^{n\beta\xi(J_pk-\mu-J_gv'(k-k^\star))}$$
$$- \frac{1}{\beta n}\log\int \mathrm{d}x\,\mathrm{d}\xi\; p(\xi)\Psi(x-J_pm\xi)\int \mathrm{d}\eta\; w(\eta)\,\mathrm{e}^{n\beta[B(x,\eta J_s)-v\eta]}. \tag{54}$$

4. Solution of order parameter equations for special cases

4.1. The state without the secondary structure

Our equations always allow for solutions with $m = 0$, describing states where no secondary structure develops. To see this we first note that now $\Psi(x) = \Phi(x)$ for any x, that (36) and (37) obey $A(-x, y) = -A(x, y)$ and $B(-x, y) = B(x, y)$, and that

$$\Lambda_\Psi(x, x'|0, k) = \left\langle \delta[x - A(x', \eta J_s)]\,\mathrm{e}^{n\beta[B(x',\eta J_s)-v\eta]}\right\rangle_\eta \left\langle \mathrm{e}^{n\beta\xi(J_pk-\mu-J_gv'(k-k^\star))}\right\rangle_\xi. \tag{55}$$

Due to the above symmetries of $A(x, y)$ and $B(x, y)$, one has $\Lambda_\Psi(x, x') = \Lambda_\Psi(-x, -x')$, so Λ_Ψ commutes with the parity operator. Its eigenfunctions are therefore either symmetric or anti-symmetric. The anti-symmetric eigenfunctions are ruled out by the requirement $\Psi(x) \geqslant 0$, so we conclude that $\Psi(x)$ must be symmetric in x, and that therefore $W(-h, \xi) = W(h, \xi)$. From this it follows, via the saddle-point equation for m, that $m = 0$ indeed solves our equations for any choice of the control parameters.

The distribution $\Psi(x)$ is for $m = 0$ to be solved from

$$\Psi(x) = \frac{\int \mathrm{d}x'\,\Psi(x')\int \mathrm{d}\eta\; w(\eta)\delta[x - A(x', \eta J_s)]\,\mathrm{e}^{n\beta[B(x',\eta J_s)-v\eta]}}{\int \mathrm{d}x'\,\Psi(x')\int \mathrm{d}\eta\; w(\eta)\,\mathrm{e}^{n\beta[B(x',\eta J_s)-v\eta]}}. \tag{56}$$

This equation has the trivial solution $\Psi(x) = \delta(x)$, which is in fact unique. To prove uniqueness we use $|\tanh[\beta A(x', \eta J_s)]| = |\tanh(\beta x')|\tanh(\beta J_s)$. Since $\Psi(x) = 0$ for $|x| > J_s$ we can define the largest interval $[-u, u] \subseteq [-J_s, J_s]$ such that $\Psi(x) = 0$ for $x \notin [-u, u]$. Inside the numerator of (56) we now know that any nonzero contribution to the integral must have $|x'| \leqslant u$, so $|\tanh[\beta A(x', \eta J_s)]| \leqslant |\tanh(\beta u)|\tanh(\beta J_s)$. Hence equation (56) tells us that if $\Psi(x) \neq 0$ then $|\tanh(\beta x)| \leqslant \tanh(\beta J_s)|\tanh(\beta u)|$, but now one must also have $|\tanh(\beta u)| \leqslant \tanh(\beta J_s)|\tanh(\beta u)|$. Clearly the only u that satisfies the latter inequality is $u = 0$, which completes the proof that $\Psi(x) = \delta(x)$.

Furthermore, upon inserting $\Psi(x) = \delta(x)$ equation (52) tells us that $W(h, \xi) = p(\xi)\delta(h)$, with $p(\xi)$ given by (45). This makes sense, since the contribution to the fields that depends on the polarity does so via the mean-field forces, which are absent for $m = 0$, whereas in the absence of long-range folding forces the remaining one-dimensional chain cannot order (hence all effective fields are zero). The saddle-point equation for k can be simplified to

$$k = \frac{\int \mathrm{d}\xi\;\xi w(\xi)\,\mathrm{e}^{n\beta\xi[J_pk-\mu-J_gv'(k-k^\star)]}}{\int \mathrm{d}\xi\; w(\xi)\,\mathrm{e}^{n\beta\xi[J_pk-\mu-J_gv'(k-k^\star)]}}. \tag{57}$$

This equation shows that even for $m = 0$ (i.e. no secondary structure) there is still an effect of the coupling between sequence selection and residue orientation: there will still be an effective preference for homogeneous sequences, due to the increased potential for energy gain (via J_p) if monomers are of the same type, which is however counter-acted by the energy cost of polarity homogeneity as controlled by J_g.

Finally, using the above results as well as $B(0, \eta J_s) = \beta^{-1} \log[2\cosh(\beta v J_s)]$, the free energy of the $m = 0$ state is seen to take the value

$$\varphi = \frac{1}{2} J_p k^2 + J_g[v(k-k^\star) - kv'(k-k^\star)] - \frac{1}{\beta n} \log \int \mathrm{d}\xi\, w(\xi)\, \mathrm{e}^{n\beta\xi[J_p k - \mu - J_g v'(k-k^\star)]}$$
$$- \frac{\log 2}{\beta} - \frac{\log \Lambda}{\beta n} - \frac{1}{\beta n} \log \int \mathrm{d}\eta\, w(\eta)\, \mathrm{e}^{n[\log\cosh(\beta v J_s) - \beta v \eta]}. \tag{58}$$

4.2. Analytical solution in verifiable limits

4.2.1. Infinite temperature. In the infinite temperature limit $\beta \to 0$ one has $A(x, y) = \beta xy + \mathcal{O}(\beta^3)$ and $B(x, y) = \beta^{-1} \log 2 + \frac{1}{2}\beta(x^2 + y^2) + \mathcal{O}(\beta^3)$. From this we can immediately extract the following solution of our saddle-point equations (47), (50) and (51):

$$\lim_{\beta\to 0} \Psi(x) = \delta(x), \qquad \lim_{\beta\to 0} W(\xi, h) = w(\xi)\delta(h), \tag{59}$$

$$\lim_{\beta\to 0} m = 0, \qquad \lim_{\beta\to 0} k = \int \mathrm{d}\xi\, \xi w(\xi). \tag{60}$$

The corresponding value for the free energy (54) is

$$\lim_{\beta\to 0} \beta f = -n^{-1} \log \Lambda - \log 2. \tag{61}$$

This is just the $\beta \to 0$ limit of the $m = 0$ (swollen) state. We recognize the free energy reducing to the entropic contributions from the angular $(-\beta^{-1} \log 2)$ and from the sequence $(-(\beta n)^{-1} \log \Lambda)$ degrees of freedom, and the average polarity k reduces to that of the amino-acid pool. All this is easily understood on physical grounds.

4.2.2. Random sequences: $n \to 0$. According to $n = \tilde{\beta}/\beta$ this limit describes the case where monomer sequences are selected fully randomly, independent of the functionality potential or the secondary structure they would generate. Our equations must for $n \to 0$ therefore reproduce the theory developed for random hetero-polymers in [26], provided we set the hydrogen bond coupling in [26] to zero. Here we find for $n \to 0$ that our equations indeed simplify considerably. As expected we obtain $p(\xi) = w(\xi)$, since sequences are selected randomly from the amino-acid pool, and hence $k = \langle \xi \rangle_\xi$. We are then left with the following eigenvalue problem for $\Psi(x)$:

$$\Psi(x) = \int \mathrm{d}y\, \Psi(y) \langle\!\langle \delta[x - A(y + J_p m\xi, \eta J_s)] \rangle\!\rangle_{\xi,\eta} \tag{62}$$

with $\Phi(x) = \langle \Psi(x - J_p m\xi) \rangle_\xi$. But now the order parameter m (which measures the degree of orientation specificity along the chain of hydrophobic versus hydrophilic residues) is to be solved from

$$m = \left\langle \xi \int \mathrm{d}h\, W(h|\xi) \tanh[\beta h] \right\rangle_\xi, \tag{63}$$

$$W(h|\xi) = \int \mathrm{d}x\, \Psi(x)\Psi(h - x - J_p m\xi). \tag{64}$$

Equivalently, upon using (62):

$$m = \int \mathrm{d}x\, \mathrm{d}x'\, \Phi(x')\Psi(x) \langle \langle \xi \tanh[\beta(x + \xi J_p m + A(x', \eta J_s))] \rangle \rangle_{\xi,\eta}. \tag{65}$$

The corresponding free energy per monomer is

$$\lim_{n\to 0}\left(f+\frac{\log\Lambda}{\beta n}\right)=\frac{1}{2}J_p\left(m^2-\langle\xi\rangle_\xi^2\right)+\mu\langle\xi\rangle+\nu\langle\eta\rangle_\eta+J_g v(\langle\xi\rangle_\xi-k^\star)$$
$$-\int \mathrm{d}x\,\Phi(x)\langle B(x,\eta J_s)\rangle_\eta. \tag{66}$$

The theory for random sequences in [26] (based on random field techniques rather than the replica formalism) involved as its main order parameter the distribution $P_\infty(\mathbf{k}|\beta J_p m)$ of three ratios $\mathbf{k}=(k_1,k_2,k_3)$ of conditioned partition functions (condition on the values of the last two spins of the chain). The link between our present equations and those in [26] is made via the identification

$$P_\infty(\mathbf{k}|\beta J_p m)=\delta(k_3-k_1)\int \mathrm{d}x\,\mathrm{d}y\,\Phi(x)\Psi(y)\delta(k_1-\mathrm{e}^{2\beta y})\delta(k_2-\mathrm{e}^{2\beta(x-y)}). \tag{67}$$

Using $\Phi(x)=\langle\Psi(x-J_p m\xi)\rangle_\xi$, equation (62) and the relation $A(x,y)=(2\beta)^{-1}\log[\cosh(\beta x+\beta y)/\cosh(\beta x-\beta y)]$ one proves that (67) obeys

$$P_\infty(\mathbf{k}|\beta J_p m)=\int \mathrm{d}\mathbf{k}'\,P_\infty(\mathbf{k}'|\beta J_p m)\left\langle\!\!\left\langle \delta\left[\mathbf{k}-\begin{pmatrix}\mathcal{F}_1(\mathbf{k}'|\beta J_s\eta)\\ \mathcal{F}_2(\mathbf{k}'|\beta J_s\eta,\beta J_p m\xi)\\ \mathcal{F}_3(\mathbf{k}'|\beta J_s\eta)\end{pmatrix}\right]\right\rangle\!\!\right\rangle_{\xi,\eta}$$

with

$$\mathcal{F}_1(\mathbf{k}|x)=\frac{\mathrm{e}^x k_1k_2+\mathrm{e}^{-x}}{\mathrm{e}^{-x}k_1k_2+\mathrm{e}^x},\qquad \mathcal{F}_2(\mathbf{k}|x,y)=\frac{\mathrm{e}^{-x}k_1k_2+\mathrm{e}^x}{\mathrm{e}^x k_1k_2+\mathrm{e}^{-x}}k_3\mathrm{e}^{2y}, \tag{68}$$

$$\mathcal{F}_3(\mathbf{k}|x)=\frac{\mathrm{e}^x k_2k_3+\mathrm{e}^{-x}}{\mathrm{e}^{-x}k_2k_3+\mathrm{e}^x} \tag{69}$$

which is indeed the limit $J_{Hb}\to 0$ of the equation derived for $q=2$ in [26].

4.2.3. Mean-field limit. A second limit which can be verified using earlier work is that where $J_s\to 0$ and $J_g\to 0$, describing the coupled dynamics of sequence selection and secondary structure generation in heteropolymers with (one type of) polarity energies only, the simpler case studied in [35]. In this limit the model contains only mean-field forces, and no longer involves transfer matrices. Using the identities $A(x,0)=0$ and $B(x,0)=\beta^{-1}\log[2\cosh(\beta x)]$ one extracts from (47) that $\Psi(x)=\delta(x)$, and so

$$m=\int \mathrm{d}\xi\,\mathrm{d}h\,W(h,\xi)\xi\tanh[\beta h], \tag{70}$$

$$k=\int \mathrm{d}\xi\,\mathrm{d}h\,W(h,\xi)\xi, \tag{71}$$

$$W(h,\xi)=\frac{\cosh^n[\beta h]p(\xi)\delta[h-\xi J_p m]}{\int \mathrm{d}h'\cosh^n[\beta h']\int \mathrm{d}\xi' p(\xi')\delta[h'-\xi' J_p m]}. \tag{72}$$

As could have been expected, the equations for the scalar order parameters (m,k) already close onto themselves. In explicit form, they are

$$m=\frac{\langle\xi\tanh(\beta\xi J_p m)\,\mathrm{e}^{n\beta\xi(J_p k-\mu)}\cosh^n(\beta\xi J_p m)\rangle_\xi}{\langle\mathrm{e}^{n\beta\xi(J_p k-\mu)}\cosh^n(\beta\xi J_p m)\rangle_\xi}, \tag{73}$$

$$k = \frac{\langle \xi \, \mathrm{e}^{n\beta\xi(J_p k-\mu)} \cosh^n(\beta\xi J_p m)\rangle_\xi}{\langle \mathrm{e}^{n\beta\xi(J_p k-\mu)} \cosh^n(\beta\xi J_p m)\rangle_\xi}. \tag{74}$$

The amino-acid statistics in equilibrium are given by

$$p(\xi) = \frac{w(\xi)\,\mathrm{e}^{n\beta\xi(J_p k-\mu)}}{\int \mathrm{d}\xi' w(\xi')\,\mathrm{e}^{n\beta\xi'(J_p k-\mu)}}, \qquad p(\eta) = \frac{w(\eta)\,\mathrm{e}^{-n\beta\nu\eta}}{\int \mathrm{d}\eta' w(\eta')\,\mathrm{e}^{-n\beta\nu\eta'}}. \tag{75}$$

Finally, using $\Phi(x) = \int \mathrm{d}\xi\; p(\xi)\delta[x - \xi J_p m]$ we may work out the value of the free energy per monomer for $J_s = 0$:

$$\lim_{J_s\to 0} f = \frac{1}{2}J_p(m^2+k^2) - \frac{\log\Lambda}{\beta n} - \frac{\log 2}{\beta} - \frac{1}{\beta n}\log\int \mathrm{d}\eta\, w(\eta)\,\mathrm{e}^{-n\beta\nu\eta}$$
$$- \frac{1}{\beta n}\log\int \mathrm{d}\xi\; w(\xi)\cosh^n(\beta\xi J_p m)\,\mathrm{e}^{n\beta\xi(J_p k-\mu)}. \tag{76}$$

The specific model studied in [35] had $\xi \in \{-1, 1\}$ and $w(\xi) = \frac{1}{2}(\delta_{\xi,1} + \delta_{\xi,-1})$, i.e. no varying degrees of hydrophobicity or hydrophilicity and a polarity-unbiased amino-acid pool. Steric effects did not come into play in [35] (monomers were characterized only by their polarity), so we may here simply take $\nu = 0$. These choices simplify our two remaining order parameter equations (73) and (74) to

$$m = \tanh(\beta J_p m), \qquad k = \tanh[n\beta(J_p k - \mu)]. \tag{77}$$

These equations are indeed identical to those of [35], given $q = 2$. Similarly, for $w(\xi) = \frac{1}{2}(\delta_{\xi,1} + \delta_{\xi,-1})$ and $\nu = 0$ the free energy per monomer (76) now simplifies to

$$\lim_{J_s\to 0} f = \frac{1}{2}J_p m^2 - \frac{\log(\Lambda/2)}{\beta n} - \frac{1}{\beta}\log[2\cosh(\beta J_p m)]$$
$$+ \frac{1}{2}J_p k^2 - \frac{1}{\beta n}\log[2\cosh(n\beta(J_p k - \mu))]. \tag{78}$$

Apart from the excess entropy $-\log(\Lambda/2)/\beta n$ due to the extra chemical degrees of freedom of our present monomers compared to those in [35] (and modulo a trivial typo in [35]) this is indeed the free-energy expression found in [35] for $q = 2$.

5. Transitions and phase diagrams for deterministic sequence selection

We now turn to non-trivial regimes where an analytical solution is still possible, but where our model does not map onto any existing model in the literature. The biologically most relevant regime is that of low or even absent genetic noise levels, namely $n \to \infty$. We still have to select a form for the polarity balance potential. Since $v(k - k^\star)$ must be minimal at $k = k^\star$ and increase monotonically with $|k - k^\star|$, we choose a simple quadratic form $v(u) = \frac{1}{2}u^2$. Thus from now on we will have $v'(u) = u$ and $v''(u) = 1$.

Since $n = \tilde{\beta}/\beta$, the limit $n \to \infty$ corresponds to the case where monomer sequences are selected fully deterministically, such as to minimize the effective Hamiltonian (5). Here, in view of many exponents in our equations growing with n, we may evaluate virtually all integrations by steepest descent. With a modest amount of foresight we define the canonical polarity balance k_0 as

$$k_0 = \frac{k^\star - \mu/J_g}{1 - J_p/J_g}. \tag{79}$$

Clearly $\lim_{J_g\to\infty} k_0 = k^\star$ and $\lim_{J_g\to 0} k_0 = \mu/J_p$. To keep our analysis as transparent as possible we will not consider pathological parameter coincidences but restrict our discussion

to the generic scenario where $J_g \neq J_p$, $\nu \neq 0$, and $k_0 \in (-1, 1)$; the system behavior in the pathological cases can always be understood as specific limits and/or degeneracies of the more generic solutions. Here, the order parameters $\{m, k, \Psi(x)\}$ are to be found by analyzing the solutions for $n \to \infty$ of the following equations, where the complications are mainly in the subtle dependence of the distribution $\Psi(x)$ on n:

$$\Psi(x) = \frac{\int \mathrm{d}x' \int \mathrm{d}\xi\ p(\xi)\Psi(x') \int \mathrm{d}\eta\ w(\eta)\delta[x - A(x' + J_p m\xi, \eta J_s)]\, \mathrm{e}^{n\beta[B(x'+J_p m\xi, \eta J_s)-\nu\eta]}}{\int \mathrm{d}x' \int \mathrm{d}\xi\ p(\xi)\Psi(x') \int \mathrm{d}\eta\ w(\eta)\, \mathrm{e}^{n\beta[B(x'+J_p m\xi, \eta J_s)-\nu\eta]}}, \tag{80}$$

$$m = \frac{\int \mathrm{d}\xi\ p(\xi)\xi \int \mathrm{d}x\, \mathrm{d}y\ \Psi(x)\Psi(y) \tanh[\beta(J_p m\xi + x + y)] \cosh^n[\beta(J_p m\xi + x + y)]}{\int \mathrm{d}\xi\ p(\xi) \int \mathrm{d}x\, \mathrm{d}y\ \Psi(x)\Psi(y) \cosh^n[\beta(J_p m\xi + x + y)]}, \tag{81}$$

$$k = \frac{\int \mathrm{d}\xi\ p(\xi)\xi \int \mathrm{d}x\, \mathrm{d}y\ \Psi(x)\Psi(y) \cosh^n[\beta(J_p m\xi + x + y)]}{\int \mathrm{d}\xi\ p(\xi) \int \mathrm{d}x\, \mathrm{d}y\ \Psi(x)\Psi(y) \cosh^n[\beta(J_p m\xi + x + y)]}, \tag{82}$$

with the abbreviations

$$p(\xi) = \frac{w(\xi)\, \mathrm{e}^{n\beta\xi(J_p - J_g)(k-k_0)}}{\int \mathrm{d}\xi' w(\xi')\, \mathrm{e}^{n\beta\xi'(J_p - J_g)(k-k_0)}}, \tag{83}$$

$$k_0 = (k^\star - \mu/J_g)/(1 - J_p/J_g). \tag{84}$$

Once the above equations have been solved for $n \to \infty$, the associated values of the free energy per monomer subsequently follows upon taking the $n \to \infty$ limit in (54).

5.1. The two simple cases $J_s = 0$ and $J_p m = 0$

In both these special cases our problem simplifies significantly due to $\Psi(x) = \delta(x)$ (a property which has been established earlier). If we define

$$L(\xi) = \frac{1}{\beta} \log \cosh(\beta J_p m\xi) + \xi(J_p - J_g)(k - k_0), \tag{85}$$

we see that our remaining equations for m and k reduce to a simple form in which for $n \to \infty$ the integration over ξ is dominated by the maximum of $L(\xi)$, subject to the constraint $\xi \in [-1, 1]$ imposed by the measure $w(\xi)$:

$$m = \lim_{n\to\infty} \frac{\int \mathrm{d}\xi\ w(\xi)\xi \tanh(\beta J_p m\xi)\, \mathrm{e}^{n\beta L(\xi)}}{\int \mathrm{d}\xi\ w(\xi)\, \mathrm{e}^{n\beta L(\xi)}}, \tag{86}$$

$$k = \lim_{n\to\infty} \frac{\int \mathrm{d}\xi\ w(\xi)\xi\, \mathrm{e}^{n\beta L(\xi)}}{\int \mathrm{d}\xi\ w(\xi)\, \mathrm{e}^{n\beta L(\xi)}}. \tag{87}$$

If $J_p m \neq 0$ then $L(\xi)$ is maximal either for $\xi = \mathrm{sgn}[(J_p - J_g)(k - k_0)]$ (if $\lim_{n\to\infty} k \neq k_0$), or for $\xi = \pm 1$ (if $\lim_{n\to\infty} k = k_0$). In either case one has $\xi \in \{-1, 1\}$, so we always find for $n \to \infty$ the simple Curie–Weiss law $m = \tanh(\beta J_p m)$ which describes a transition to secondary structure at $T = J_p$. Our equation for k, on the other hand, will produce for $n \to \infty$ only solutions of

$$k = \mathrm{sgn}[(J_p - J_g)(k - k_0)] \tag{88}$$

(this includes the case $k = k_0$). Graphical inspection of this equation shows immediately that for $J_p < J_g$ the only solution is $k = k_0$, whereas for $J_p > J_g$ we have the additional

solutions $k = \pm 1$. If $J_p m = 0$ then $L(\xi)$ is either maximal for $\xi = \mathrm{sgn}[(J_p - J_g)(k - k_0)]$ (if $\lim_{n\to\infty} k \neq k_0$), or it is a constant on $\xi \in [-1, 1]$ (if $\lim_{n\to\infty} k = k_0$). Here one has $\xi \in \{-1, 1\}$ only if $k \neq k_0$.

Working out the free energy per monomer (54) gives, using $B(x, y) = B(|x|, |y|)$ and the property that $B(|x|, |y|)$ increases monotonically with both $|x|$ and $|y|$,

$$\begin{aligned}
\varphi &= \frac{1}{2}J_p m^2 + \frac{1}{2}J_g k^{\star 2} + \frac{1}{2}(J_p - J_g)k^2 \\
&\quad - \lim_{n\to\infty} \frac{1}{\beta n} \log \int \mathrm{d}\eta\, w(\eta) \int \mathrm{d}\xi\, w(\xi)\, \mathrm{e}^{n\beta[\xi(J_p-J_g)(k-k_0)+B(J_p m\xi,\eta J_s)-\nu\eta]}, \\
&= \frac{1}{2}J_p m^2 + \frac{1}{2}J_g k^{\star 2} + \frac{1}{2}(J_p - J_g)k^2 \\
&\quad - \max_{\xi,\eta\in[-1,1]} \{\xi(J_p - J_g)(k - k_0) + B(J_p m\xi, \eta J_s) - \nu\eta\}, \\
&= \frac{1}{2}J_p m^2 - B(J_p|m|, J_s) + \frac{1}{2}J_g k^{\star 2} - |\nu| + \frac{1}{2}(J_p - J_g)k^2 - |J_p - J_g||k - k_0|, \qquad (89)
\end{aligned}$$

where the maximum corresponds to $\eta = -\mathrm{sgn}(\nu)$ and $\xi = \mathrm{sgn}[(J_p - J_g)(k - k_0)]$. The last line reveals that in cases where we have multiple solutions, namely $J_p > J_g$, the solution $k = k_0$ is always a local maximum of φ and $k = \pm 1$ are always local minima. Of the latter two, the lowest free energy is found for $k = -\mathrm{sgn}(k_0)$ (this is therefore the state that is not only locally stable but also thermodynamically stable). Therefore

$$J_p > J_g: \quad k = \pm 1, \qquad J_p < J_g: \quad k = k_0 \qquad (90)$$

This implies a discontinuous phase transition at $J_p = J_g$, where we go from $k = \pm 1$ (homogeneous polarity sequences) to $k = k_0 \in (-1, 1)$, where the sequence becomes inhomogeneous in polarity.

If we calculate the distribution $W(\xi, h)$ for the above solutions we always find $W(\xi, h) = \pi(\xi)\delta(h)$, but with potentially different polarity statistics. For the $k = \pm 1$ states one has $\pi(\xi) = \delta(\xi - k)$. For the $k = k_0$ solution, however, we need to look beyond the leading order and write $k = k_0 + n^{-1}k_1 + \mathcal{O}(n^{-2})$. Here we find for $n \to \infty$:

$$\pi(\xi) = \frac{w(\xi)\, \mathrm{e}^{\beta\xi(J_p-J_g)k_1}}{\int \mathrm{d}\xi' w(\xi')\, \mathrm{e}^{\beta\xi'(J_p-J_g)k_1}} \qquad (91)$$

with k_1 to be solved from

$$k_0 = \frac{\int \mathrm{d}\xi\, w(\xi)\xi\, \mathrm{e}^{\beta\xi(J_p-J_g)k_1}}{\int \mathrm{d}\xi\, w(\xi)\, \mathrm{e}^{\beta\xi(J_p-J_g)k_1}}. \qquad (92)$$

This concludes our solution for the simple cases $J_s = 0$ and $J_p m = 0$. From now on we consider the case where $J_s > 0$ and $J_p m \neq 0$.

5.2. Summary of the $n \to \infty$ theory

The full analysis of our order parameter equations in the limit $n \to \infty$ via saddle-point analysis, for arbitrary (J_s, J_p), turns out to be non-trivial; details of this calculation would interrupt the flow of the paper and have therefore been delegated to appendix B. The end result, however, is surprisingly simple. We can summarize the final equations for our order parameters (k, m) describing the system states as identified in the limit $n \to \infty$ in the following compact way:

$$J_g > J_p: \qquad k = k_0, \qquad m = 0 \quad \text{or} \quad F_{\beta J_p}(m) = -\tanh(\beta J_s), \qquad (93)$$

$$J_p > J_g: \qquad k = \pm 1, \qquad m = 0 \quad \text{or} \quad F_{\beta J_p}(m) = \mathrm{sgn}(\nu)\tanh(\beta J_s), \qquad (94)$$

in which the function $F_x(m)$ is defined as

$$F_x(m) = \frac{\tanh\left[\frac{1}{2}xm - \frac{1}{2}\tanh^{-1}(m)\right]}{\tanh\left[\frac{1}{2}xm + \frac{1}{2}\tanh^{-1}(m)\right]}. \tag{95}$$

Only the solutions with $k = k_0$ as obtained for $J_g > J_p$ correspond to hetero-polymers with inhomogeneous polarity along the chain, i.e. to systems of the protein type. The solutions with $k = \pm 1$ (with $k = -\mathrm{sgn}(k_0)$ being also thermodynamically stable) describe a situation where the sequence selection results in polymers with homogeneous polarity. For $J_p > J_g$ we have two further conditions (B.26) and (B.27); these are always satisfied for $m = 0$, but may be violated by saddle points for which $|m|$ is too large. We observe that for $\nu < 0$ the homogeneous polarity states and the inhomogeneous polarity states exhibit fully identical levels of the secondary structure (as measured by m), for any combination of βJ_p and βJ_s. Here it is therefore also easy to show by comparing the two free-energy expressions (B.29) and (B.45) that for $J_p > J_g$ the free energy per monomer of the $k = \pm 1$ state is lower than that of the state $k = k_0$, whereas for $J_g > J_p$ the free energy of the $k = k_0$ state is lower. For $\nu > 0$, however, the two states no longer have identical values of m, with that of the $k = \pm 1$ state being lower; here the system finds it increasingly difficult to combine homogeneous polarity sequences with the secondary structure.

Let us inspect the bifurcation phenomenology for the order parameter m. Note that $F_0(m) = -1$ for all $m \in [-1, 1]$, and that $F_\infty(m) = 1$ for all $m \in [-1, 1]$. For $x > 0$ the function $F_x(m)$ is symmetric in m, with $F_x(\pm 1) = -1$ and with

$$F_x(m) = \frac{x-1}{x+1} - m^2\frac{x(3-x^2)}{3(x+1)^2} + \mathcal{O}(m^4) \tag{96}$$

(see also figure 4). In view of the symmetry $F_x(-m) = F_x(m)$, we conclude that (depending on the values of (x, y)), the equation $F_x(m) = y$ has either zero, two ($\pm m^\star$), or four ($\pm m^\star, \pm m_0$) non-trivial solutions in m.

In the (x, y) plane, where $x = \beta J_p$ and $y = \tanh(\sigma\beta J_s)$ with $\sigma = \pm 1$ (so $\sigma = \mathrm{sgn}(\nu)$ for $J_p > J_g$ and $\sigma = -1$ for $J_g > J_p$), the bifurcation scenarios for our saddle-point equation $F_x(m) = y$ can now be summarized as:

$$\begin{aligned}
x < \sqrt{3}: &\quad \text{continuous transition at } y_c = (x-1)/(x+1)\\
&\quad y < y_c: \quad m \in \{0, \pm m^\star(x)\}\\
&\quad y > y_c: \quad m = 0,\\
x > \sqrt{3}: &\quad \text{continuous transition at } y_c = (x-1)/(x+1)\\
&\quad y < y_c: \quad m \in \{0, \pm m^\star(x)\}\\
&\quad y > y_c: \quad m \in \{0, \pm m_0(x), \pm m^\star(x)\}\\
&\quad \text{discontinuous transition at} y_c' > (x-1)/(x+1)\\
y < y_c': &\quad m\{0, \pm m_0(x), \pm m^\star(x)\}\\
y > y_c': &\quad m = 0.
\end{aligned}$$

The result is shown in figure 5.

5.3. Phases, transition lines and phase diagrams

We can characterize the phases of our system for $n \to \infty$ in terms of the values for the order parameters (k, m), where k provides information on the primary structure (the average polarity) and m provides information on the secondary structure (the extent of order in the

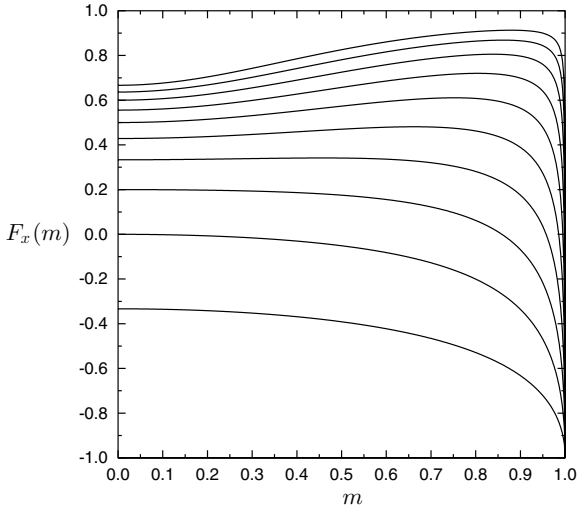

Figure 4. The function $F_x(m)$ for $x \in \{\frac{1}{2}, 1, \frac{3}{2}, 2, \frac{5}{2}, 3, \frac{7}{2}, 4, \frac{9}{2}, 5\}$ (from bottom to top). Solving the equation $F_x(m) = y$ for m can give at most one positive solution if $x < \sqrt{3}$, where $[\mathrm{d}^2 F_x(m)/\mathrm{d}m^2]|_{m=0} < 0$. It may have two positive solutions if $x > \sqrt{3}$, where $[\mathrm{d}^2 F_x(m)/\mathrm{d}m^2]|_{m=0} > 0$, provided one also has $y > F_x(0) = (x-1)/(x+1)$. For sufficiently large y the equation $F_x(m) = y$ will no longer have any solutions.

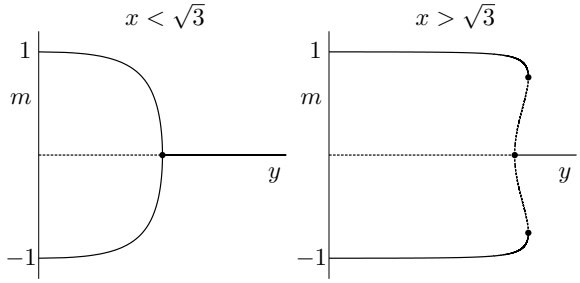

Figure 5. The bifurcation scenarios for the solutions m of the equation $F_x(m) = y$, with $x = \beta J_p$ and $y = \mathrm{sgn}(\nu)\tanh(\beta J_s) \in [-1, 1]$ for $J_p > J_g$ and $y = -\tanh(\beta J_s) \in [-1, 0]$ for $J_g > J_p$. Solid lines correspond to stable solutions (local minima of the free energy), whereas dashed lines correspond to unstable ones. The trivial solution $m = 0$ changes stability at $\beta J_s = \frac{1}{2}\mathrm{sgn}(\nu)\log(\beta J_p)$ for $J_p > J_g$ and at $\beta J_s = \frac{1}{2}\log(\beta J_p)$ for $J_g > J_p$.

side-chain orientations). The system is found to exhibit five phases:

HS (homogeneous and swollen): $k = \pm 1$, $m = 0$
primary structure but no secondary structure,
selected sequences are homogeneous in polarity.

HC (homogeneous and collapsed): $k = \pm 1$, $m \neq 0$
both primary and secondary structures,
selected sequences are homogeneous in polarity.

HM (homogeneous and mixed): $k = \pm 1$, coexistence of $m = 0$ and $m \neq 0$
primary structure, with secondary structure controlled by remanence,
sequences are homogeneous in polarity.

IS (inhomogeneous and swollen): $k = k_0, m = 0$
primary structure but no secondary structure,
selected sequences are inhomogeneous in polarity.

IC (inhomogeneous and collapsed): $k = k_0, m \neq 0$
both primary and secondary structures,
selected sequences are inhomogeneous in polarity.

There is no random (paramagnetic) phase $m = k = 0$. This is a consequence of the $n \to \infty$ limit: since the noise in the genetic selection (representing mutations) is removed, there is at least always a primary structure developing as measured by $k \neq 0$.

Similarly, we can summarize the transitions we have by now identified:

- HS $\to$ IS and HC $\to$ IC: discontinuous transitions, at
$$J_g = J_p. \tag{97}$$
The HS $\to$ IS line is found in the regime of small values of J_p. The HC $\to$ IC line is found for large values of J_p. Along the latter line, if $\nu < 0$ only k is changed at the transition, if $\nu > 0$ both k and m are changed.
- HS $\to$ HC, IS $\to$ IC and HC $\to$ HM: continuous transitions, at
$$\beta J_s = \begin{cases} \frac{1}{2}\mathrm{sgn}(\nu)\log(\beta J_p) & \text{if} \quad J_p > J_g \\ -\frac{1}{2}\log(\beta J_p) & \text{if} \quad J_g > J_p. \end{cases} \tag{98}$$
The HC $\to$ HM line exists only when $J_p > J_g$ and $\nu > 0$ (where the coexistence phase HM is found).
- HS $\to$ HM: discontinuous transition, to be solved from the coupled equations
$$\frac{\tanh\left[\frac{1}{2}\beta J_p m - \frac{1}{2}\tanh^{-1}(m)\right]}{\tanh\left[\frac{1}{2}\beta J_p m + \frac{1}{2}\tanh^{-1}(m)\right]} = \tanh(\beta J_p), \tag{99}$$
$$\frac{1-\tanh^2\left[\frac{1}{2}\beta J_p m - \frac{1}{2}\tanh^{-1}(m)\right]}{1-\tanh^2\left[\frac{1}{2}\beta J_p m + \frac{1}{2}\tanh^{-1}(m)\right]}\frac{\beta J_p(1-m^2)-1}{\beta J_p(1-m^2)+1} = \tanh(\beta J_s), \tag{100}$$
where the second equation is obtained from combining $F_{\beta J_p}(m) = \tanh(\beta J_s)$ with $\frac{\partial}{\partial m}F_{\beta J_p}(m) = 0$. This line starts at the triple point $(\beta J_p, \beta J_s) = \left(\sqrt{3}, \frac{1}{4}\log 3\right)$ in the $(\beta J_p, \beta J_s)$ plane, and rises continually for $\beta J_p > \sqrt{3}$. It emerges only for $J_p > J_g$ and $\nu > 0$ (where the coexistence phase HM is found).

At the continuous transition (98) the $m \neq 0$ state always takes over the stability from the trivial one. This can be seen upon expanding the two free-energy expressions (B.29) and (B.45) for small m. For both expressions this gives
$$\beta(\varphi - \varphi_{m=0}) = \frac{1}{8}m^2(\beta J_p + 1)^2\left\{\tanh^2(\beta J_s) - \left(\frac{\beta J_p - 1}{\beta J_p + 1}\right)^2\right\} + \mathcal{O}(m^3).$$
Although both are co-located and are continuous in the fundamental order parameters (m, k), there is an important difference between the HS $\to$ HC and the IS $\to$ IC transitions, which involves the behavior of the polarity distribution $\pi(\xi)$. As one crosses from HS into HC, $\pi(\xi)$ remains unchanged, taking the value $\pi(\xi) = \delta(\xi - k)$ in both states. In contrast, we know from (91) that the IS state has a continuous polarity distribution $\pi(\xi) = \int \mathrm{d}h\, W(\xi, h)$, whereas the IC state has the binary distribution $\pi(\xi) = \frac{1}{2}(1 + k_0)\delta(\xi - 1) + \frac{1}{2}(1 - k_0)\delta(\xi + 1)$. Thus, the transition IS $\to$ IC is in fact *discontinuous*, in spite of it involving no jump in the order parameter m itself.

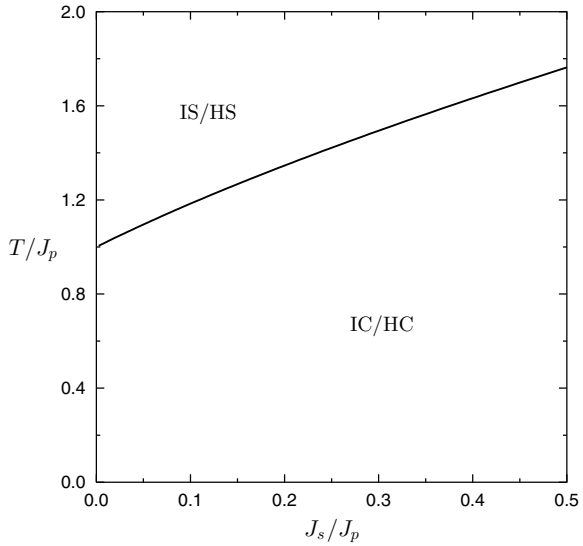

Figure 6. Phase diagram cross-section for $n \to \infty$ (deterministic sequence selection) for the cases where either $J_g > J_p$ (protein-like inhomogeneous polarity sequences, $k = k_0$) or $J_g < J_p$ (homogeneous polarity sequences, $k = \pm 1$) but with $\nu < 0$. Solid line: transition marking the continuous bifurcation of collapsed ($m \neq 0$) states, although for $J_g > J_p$ this transition is discontinuous in the polarity statistics. Phases are defined and described in the main text.

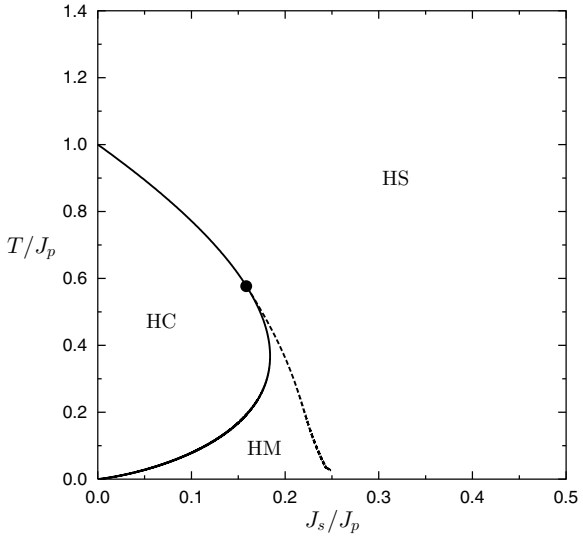

Figure 7. Phase diagram cross-section for $n \to \infty$ (deterministic sequence selection) for the case where $J_p > J_g$ and $\nu > 0$ (homogeneous polarity sequences, unlike proteins). Here the system is unable to minimize steric and polar energies simultaneously. Solid line: the continuous transitions between swollen ($m = 0$) and collapsed ($m \neq 0$) solutions. Dashed: the discontinuous transition. Phases are defined and described in the main text.

Upon translating our results into the original control parameters βJ_p and βJ_s one obtains the phase diagram cross-sections shown in figures 6 and 7. The phase where compact ($m \neq 0$)

and swollen ($m = 0$) states coexist will be characterized by strong remanence effects. The thermodynamic transition line (calculated by selecting the solution with the lowest free energy) coincides with the second-order transition for $\beta J_p < \sqrt{3}$, and will be found inside the coexistence region for $\beta J_p > \sqrt{3}$.

Without noise (i.e. random mutations) in the sequence selection process, $\tilde{\beta} = \infty$, we can summarize the behavior of the system as follows. For $J_p > J_g$ it always finds itself in states where any infinitesimal functional advantage of either the hydrophilic or the hydrophobic monomers leads to amino-acid sequences that are, unlike proteins, fully homogeneous in their polarity. The phenomenology described by the remaining equations for m and the resulting phase diagram reflect the interplay between the tendencies of the polarity-homogeneous system to have similarly oriented amino-acid residues (induced by the long-range forces) and low steric energies (induced by the short-range forces). The system behaves as an Ising chain with random short-range bonds and uniform long-range bonds. In those cases where the amino-acids are forced by steric effects to have non-identical side chain orientations (i.e. for $\nu > 0$) there is a complex competition between long-range and short-range order, which leads to low values of $|m|$ and strong remanence effects, in sharp contrast to the situation in mean-field models [35]. In contrast, for $\nu < 0$ both the long-range and the short-range forces promote similar side chain orientations; the absence of frustration is responsible for the absence of remanence effects and for having large $|m|$ (strong secondary structure). For $J_g > J_p$ it is no longer energetically advantageous to select chains with uniform polarity, and here we find the protein-like states. The polarity inhomogeneity of the sequence reduces dramatically the energetic impact of the long-range forces compared to the case $k \pm 1$, and this decouples the strength $|m|$ of the secondary structure from any preference for aligning or anti-aligning short-range forces, as controlled by ν.

6. Transitions and phase diagrams for non-deterministic sequence selection

In this section, we extract solutions, transition lines and phase diagrams from our order parameter equations for non-deterministic selection of primary sequences, namely finite n. Full analytical solution of our equations is generally ruled out, so we restrict ourselves to the study of instabilities and to collecting further information on phases by solving our equations numerically. As in the previous section we restrict ourselves to simple parameter choices, in particular we take $v(u) = \frac{1}{2}u^2$ and $k_0 \in (-1, 1)$.

6.1. Continuous transitions away from $m = 0$

We first derive exact conditions marking continuous phase transitions away from the state $m = 0$ without secondary structure as defined and studied earlier, for arbitrary n. For $m = 0$ one has $\Psi(x) = \Phi(x) = \delta(x)$, and k is to be solved from (57). We make in our order parameter equations (45), (47), (50)–(52) the substitutions $m \to \Delta m, k \to k + \Delta k$ and $\Psi(x) \to \delta(x) + \Delta\Psi(x)$. We next expand these equations in $\{\Delta m, \Delta k, \Delta\Psi(x)\}$ and locate their linear instabilities. In doing so we may use $k = \int \mathrm{d}\xi\ p(\xi)\xi$, which holds for $m = 0$. In practice it turns out somewhat easier to involve also the auxiliary distribution $\Phi(x)$, and replace (47) by the pair (48) and (49). First, substitution into and expansion of equations (45) and (49) gives

$$\Delta p(\xi) = n\beta(J_p - J_g)(\xi - k)p(\xi)\Delta k + \mathcal{O}(\Delta^2), \tag{101}$$

$$\Delta\Phi(x) = \Delta\Psi(x) - J_p k\delta'(x)\Delta m + \mathcal{O}(\Delta^2). \tag{102}$$

These results are then substituted into (48), which leads to an equation for $\Delta\Psi(x)$:

$$\Delta\Psi(x) = \frac{\int \mathrm{d}x'[\Delta\Psi(x') - J_p k \Delta m \delta'(x')] \int \mathrm{d}\eta\, w(\eta)\{\delta[x - A(x', \eta J_s)] - \delta(x)\}\, \mathrm{e}^{n\beta[B(x', \eta J_s) - \nu\eta]}}{\int \mathrm{d}\eta\, w(\eta)\, \mathrm{e}^{n\beta[B(0, \eta J_s) - \nu\eta]}} + \mathcal{O}(\Delta^2). \tag{103}$$

We next separate $\Delta\Psi(x)$ into its symmetric and anti-symmetric parts, $\Delta\Psi(x) = \Delta\Psi_S(x) + \Delta\Psi_A(x)$, giving up to order Δ:

$$\Delta\Psi_S(x) = \frac{\int \mathrm{d}x' \Delta\Psi_S(x') \int \mathrm{d}\eta\, w(\eta)\{\delta[x - A(x', \eta J_s)] - \delta(x)\}\mathrm{e}^{n\beta[B(x', \eta J_s) - \nu\eta]}}{\int \mathrm{d}\eta\, w(\eta)\, \mathrm{e}^{n\beta[B(0, \eta J_s) - \nu\eta]}}, \tag{104}$$

$$\Delta\Psi_A(x) = \frac{\int \mathrm{d}x'[\Delta\Psi_A(x') - J_p k \Delta m \delta'(x')] \int \mathrm{d}\eta\, w(\eta)\delta[x - A(x', \eta J_s)]\mathrm{e}^{n\beta[B(x', \eta J_s) - \nu\eta]}}{\int \mathrm{d}\eta\, w(\eta)\, \mathrm{e}^{n\beta[B(0, \eta J_s) - \nu\eta]}}. \tag{105}$$

The symmetric and anti-symmetric parts obey independent equations, and only the anti-symmetric part $\Psi_A(x)$ is coupled to the bifurcation of $m \neq 0$. Apparently, any nonzero solution of equation (104) describes transitions from one $m = 0$ state to another, whereas equation (105) controls the bifurcations away from $m = 0$.

In order to expand equations (50) and (51) for the scalar order parameters we need to vary the distribution $W(\xi, h)$ defined in (52), which we first rewrite as

$$W(\xi, h) = \frac{p(\xi)\cosh^n(\beta h) \int \mathrm{d}x\, \mathrm{d}y\, \Psi(x)\Psi(y)\delta(h - J_p m\xi - x - y)}{\int \mathrm{d}\xi' p(\xi') \int \mathrm{d}x\, \mathrm{d}y\, \Psi(x)\Psi(y)\cosh^n[\beta(J_p m\xi' + x + y)]}.$$

Upon varying this equation around the $m = 0$ state we then find

$$\begin{aligned}
\Delta W(\xi, h) &= \Delta p(\xi)\delta(h) + 2p(\xi)\cosh^n(\beta h)\Delta\Psi(h) - J_p \Delta m \xi p(\xi)\cosh^n(\beta h)\delta'(h) \\
&\quad - 2p(\xi)\delta(h)\int \mathrm{d}y \cosh^n(\beta y)\Delta\Psi_S(y) + \mathcal{O}(\Delta^2), \\
&= p(\xi)\Big\{n\beta(J_p - J_g)(\xi - k)\delta(h) + \cosh^n(\beta h)[2\Delta\Psi(h) - J_p \xi \Delta m \delta'(h)] \\
&\quad - 2\delta(h)\int \mathrm{d}y \cosh^n(\beta y)\Delta\Psi_S(y)\Big\} + \mathcal{O}(\Delta^2).
\end{aligned} \tag{106}$$

Insertion into (50) and (51) then gives, using $\int \mathrm{d}h \tanh(\beta h)\cosh^n(\beta h)\delta'(h) = -\beta$:

$$\Delta m = 2k\int \mathrm{d}h \tanh(\beta h)\cosh^n(\beta h)\Delta\Psi_A(h) + \beta J_p \Delta m \int \mathrm{d}\xi\; p(\xi)\xi^2 + \mathcal{O}(\Delta^2), \tag{107}$$

$$\Delta k = n\beta(J_p - J_g)\left[\int \mathrm{d}\xi\, \xi^2 p(\xi) - k^2\right]\Delta k + \mathcal{O}(\Delta^2). \tag{108}$$

As expected, the perturbations Δm couple only to the anti-symmetric part of $\Delta\Psi(x)$; the $m \neq 0$ bifurcations are the instabilities of the coupled pair (105) and (107). Furthermore, equation (108) for Δk does not depend on the symmetric part of $\Delta\Psi(x)$, so we may for the purpose of studying continuous transitions away from the $m = 0$ state regard $\delta\Psi(x)$ as strictly anti-symmetric and extract instabilities involving k only from (108).

It turns out that the (anti-symmetric) functional perturbation $\Delta\Psi_A(x)$ that solves equation (105) can be expressed in terms of Δm. We show this by substituting for $\lambda \neq 1$ the ansatz

$$\Delta\Psi_A(x) = \frac{\lambda J_p k}{\lambda - 1}\delta'(x)\Delta m, \tag{109}$$

into the leading orders of (105). Using integration by parts and the properties $\partial_x B(x, y)|_{x=0} = 0$ and $\partial_x A(x, y)|_{x=0} = \tanh(\beta y)$ this is found to give

$$\begin{aligned}\lambda\delta'(x) &= -\frac{\int dx'\delta'(x')\int d\eta\, w(\eta)\delta[x - A(x', \eta J_s)]\, e^{n\beta[B(x',\eta J_s)-\nu\eta]}}{\int d\eta\, w(\eta)\, e^{n\beta[B(0,\eta J_s)-\nu\eta]}},\\ &= -\frac{\int dx'\delta(x')\int d\eta\, w(\eta)\{n\beta\delta[x - A(x', \eta J_s)]\frac{\partial}{\partial x'}B(x', \eta J_s)\}\, e^{n\beta[B(x',\eta J_s)-\nu\eta]}}{\int d\eta\, w(\eta)\, e^{n\beta[B(0,\eta J_s)-\nu\eta]}},\\ &\quad + \frac{\int dx'\delta(x')\int d\eta\, w(\eta)\{\delta'[x - A(x', \eta J_s)]\frac{\partial}{\partial x'}A(x', \eta J_s)\}\, e^{n\beta[B(x',\eta J_s)-\nu\eta]}}{\int d\eta\, w(\eta)\, e^{n\beta[B(0,\eta J_s)-\nu\eta]}}\\ &= \delta'(x)\frac{\int d\eta\, w(\eta)\tanh(\beta\eta J_s)\, e^{n\beta[B(0,\eta J_s)-\nu\eta]}}{\int d\eta\, w(\eta)\, e^{n\beta[B(0,\eta J_s)-\nu\eta]}}. \end{aligned} \tag{110}$$

This confirms that (109) indeed solves our bifurcation equation, with

$$\lambda = \frac{\int d\eta\, w(\eta)\tanh(\beta\eta J_s)\, e^{n\beta[B(0,\eta J_s)-\nu\eta]}}{\int d\eta\, w(\eta)\, e^{n\beta[B(0,\eta J_s)-\nu\eta]}}. \tag{111}$$

This result allows us to compactify our bifurcation conditions further. Upon substituting (109) into (107) and carrying out the remaining integral, we obtain the following simple set of bifurcation conditions:

$$\Delta m \neq 0: \quad 1 = \beta J_p\left[\int d\xi\, \xi^2 p(\xi) - \frac{2\lambda k^2}{\lambda - 1}\right], \tag{112}$$

$$\Delta k \neq 0: \quad 1 = n\beta(J_p - J_g)\left[\int d\xi\, \xi^2 p(\xi) - k^2\right], \tag{113}$$

where

$$p(\xi) = \frac{w(\xi)\, e^{n\beta\xi(J_p-J_g)(k-k_0)}}{\int d\xi' w(\xi')\, e^{n\beta\xi'(J_p-J_g)(k-k_0)}}. \tag{114}$$

For $\beta = 0$, infinite temperature, the right-hand sides of (112) and (113) are zero. Hence the physical transitions occur at the highest temperature for which the right-hand sides have increased to the value 1. If the first transition to take place is (113), then m will remain zero and equation (112) will still apply to predict a further $m \neq 0$ transition. If (112) is the first transition to occur, then (113) will no longer apply.

As a simple but non-trivial test we can recover from (112) and (113) our earlier predictions for the limit $n \to \infty$. Taking $n \to \infty$ in (111) gives the simple result $\lim_{n\to\infty}\lambda = -\mathrm{sgn}(\nu)\tanh(\beta J_s)$. In the HS, HC and HM phases we have $J_p > J_g$ and $k = \pm 1$, so $\lim_{n\to\infty} p(\xi) = \delta(\xi - k)$ and therefore $\lim_{n\to\infty}\int d\xi\, \xi^2 p(\xi) = 1$. This simplifies condition (112) for the continuous bifurcation of $m \neq 0$ in the $k = \pm 1$ phases to the expression found earlier in analyzing the $n \to \infty$ equations, as it should:

$$\beta J_s = \tfrac{1}{2}\mathrm{sgn}(\nu)\log(\beta J_p). \tag{115}$$

For the $k = k_0$ states the $m \neq 0$ bifurcation is discontinuous, involving a jump in the polarity statistics as measured by $\pi(\xi)$; so there equations (112) and (113) do not apply.

As an application of (112) and (113) we have solved these equations numerically for $J_g/J_p = \nu = \frac{1}{2}$ and $k_0 = 0$, to investigate the effect of genetic noise on the phase diagram in figure 7 (although this is the biologically less relevant case of polymers with homogeneous polarity, it has the more interesting phase diagram). The result is shown in figure 8. Although based on equations that only apply to continuous transitions, the figure allows us to predict on

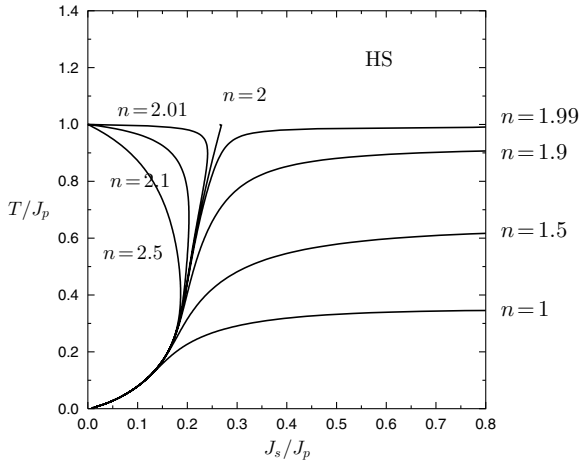

Figure 8. Continuous bifurcations from swollen ($m = 0$, HS) to collapsed ($m \neq 0$) states, for several n values around $n = 2$, for the case where $J_p > J_g$ and $\nu > 0$ (homogeneous polarity sequences, unlike proteins). The corresponding curve for $n = \infty$ is shown in figure 7. We see that, if there were no discontinuous transitions, reentrance would occur upon lowering T for $n > 2$, where for $n < 2$ the continuous transition temperature is monotonic in J_s/J_p. This suggests strongly that there is a discontinuous bifurcation to an HM phase for $n > 2$, but not for $n < 2$.

topological grounds that discontinuous transitions will occur for $n \geqslant 2$. This is a remarkable result: the critical value $n = 2$ for the onset of first-order transitions was found persistently in earlier coupled dynamics models [31–34], but since these did not involve short-range forces, its re-appearance in the present model strongly suggests an unexpected universality which at present we do not understand.

7. Numerical results

7.1. Numerical solution of order parameter equations via population dynamics

The goal of this section is to verify numerically the phases predicted in previous sections, and to provide phase diagrams for those cases where solutions of equations (47), (50)–(52) for the observables m, k and $\Psi(x)$ cannot be found analytically. To limit the number of control parameters to be varied we choose $J_s = 0.1$, $J_p = 1$, $\mu = J_p k^\star$ (so $k_0 = k^\star$) and $k^\star = 0.7$ throughout, since this still allows us to probe all the phases in figures 6 and 7. We followed the mathematically related studies [36–39] and solved the functional equation (47) using a so-called population dynamics algorithm (with a population of size 10^4), which exploits the interpretation of such equations as fixed-point conditions for a suitably chosen stochastic process for the local fields.

We turn first to the most important and realistic case of (near-)deterministic sequence selection, where for $n \to \infty$ we expect to recover the behavior shown in the phase diagrams of figures 6 and 7. Here we face the practical problem that in our equations n appears usually in exponents, which limits our numerical analysis to values $n \leqslant 400$. It turns out that to observe the $n \to \infty$ predictions one needs values of n that are significantly larger than this; furthermore, for large but finite n the limiting values of transition temperatures and the nature of the various transitions can vary significantly from one phase to another. In figure 9 we

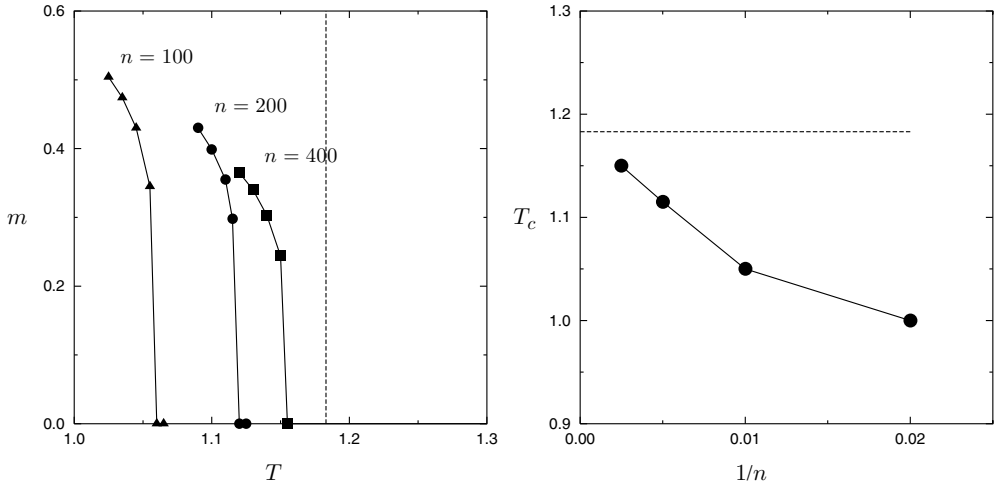

Figure 9. Left: dependence of order parameter m on the folding temperature T, obtained by numerical solution of the order parameter equations, for control parameters $(J_s, J_p, J_g) = (0.1, 1, 2), k_0 = 0.7, \mu = 0.2$ and $\nu = 0.5$. The relative genetic noise levels $n = \tilde{T}/T$ where $n = 100$ (connected triangle), $n = 200$ (connected circles) and $n = 400$ (connected squares). According to our earlier analysis, for $n \to \infty$ the phases should be those shown in figure 7. For the present values of control parameters this predicts for $n \to \infty$ a continuous transition from $m \neq 0$ (IS phase) to $m = 0$ (IC phase) at $\lim_{n\to\infty} T_c = 1.183$ (shown as a vertical dashed line). Right: the IS $\to$ IC transition temperatures T_c shown versus $1/n$, for the same values of the remaining control parameters. The data are perfectly consistent with the analytically determined value $\lim_{n\to\infty} T_c = 1.183$ (dashed).

present numerical results for positive ν, where the steric forces make it energetically favorable for adjacent amino acids to have different side chain orientations. We plot the order parameter m versus temperature (the left panel) to locate the IS $\to$ IC phase transition, which for $n \to \infty$ was predicted to be continuous, and which for the present control parameters should occur at $T_c = 1.183$. It turns out that for large but finite n the transition is in fact *discontinuous* and at a lower temperature than the $n \to \infty$ one. However, a study of the asymptotic scaling with n of the transition temperature, within the numerically accessible regime, confirms that for $n \to \infty$ the correct value is found, see figure 9 (the right panel). The observed strong dependence on n of the exact location of the transition is remarkable; the system appears to be very sensitive to the ratio of temperatures of the two coupled processes, and the deterministic regime is achieved only asymptotically. For $n = 100$ the location of the transition point differs by more than 10% from its $n \to \infty$ value. If one carries out a scaling analysis of the magnitude of the jump in m found at the transition temperature for large but finite n, one finds that for $n \to \infty$ this jump will indeed vanish, in agreement with our previous asymptotic analysis.

Upon carrying out a similar analysis for negative values of ν, where steric forces are such that adjacent amino acids prefer identical chain orientations, the resulting graphs and the physical picture are similar to those of $\nu > 0$. For large but finite n the phase transition is again discontinuous, and a scaling analysis shows once more good agreement with the theory in the limit $n \to \infty$. However, there is an important difference between the cases $\nu > 0$ and $\nu < 0$ which concerns the sub-leading orders in n^{-1} for the state $k = k_0$, as $n \to \infty$, which is reflected in both the field distribution $\Psi(x)$ and in the order parameter k close the transition. This is an important success of the population dynamics algorithm, which allows

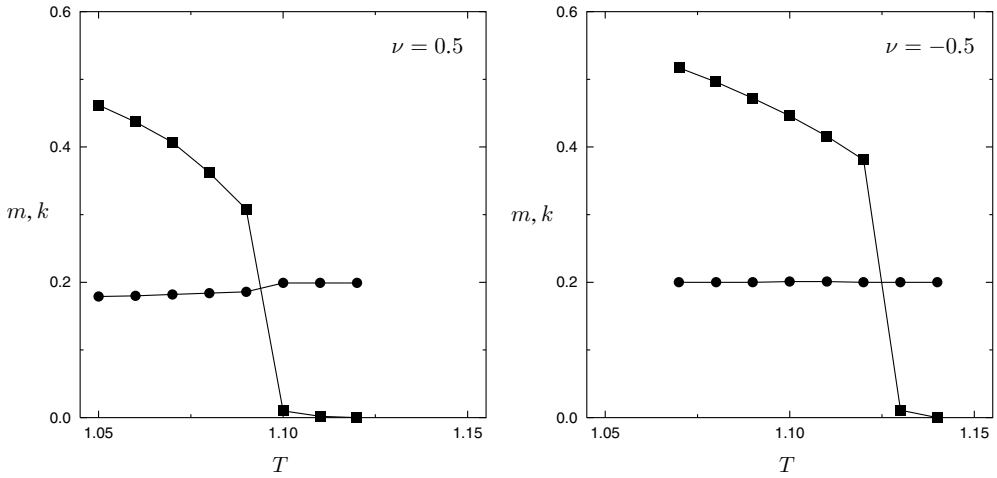

Figure 10. Dependence of order parameter m (connected squares) and k (connected circles) on the folding temperature T, obtained by numerical solution of the order parameter equations, for control parameters $(J_s, J_p, J_g) = (0.1, 1, 2)$, $k_0 = 0.2$ and $\mu = 0.7$. In both graphs the relative genetic noise level is $n = \tilde{T}/T = 200$. Left graph: $\nu = 0.5$ (promoting different orientations for adjacent amino acids). Right graph: $\nu = -0.5$ (promoting identical orientations). The large n theory of the previous section predicted that the sub-leading order in n for the $k = k_0$ solution (as shown here) is $\mathcal{O}(n^{-1})$ when $\nu < 0$, but $\mathcal{O}(n^{-1/2})$ when $\nu > 0$. The numerical data shown here are consistent with this prediction.

us to evaluate in a simple and straightforward way the distribution $\Psi(x)$ of the short-range contributions to the local effective fields. In addition, it is a crucial test to verify the scaling of the sub-leading orders in n^{-1} predicted by our theory. Figure 10 shows how m and k behave close to the transition. One notes that the (discontinuous) behavior of m is qualitatively similar in both cases, whereas the polarity k behaves in a very different way: in contrast to $\nu < 0$, for $\nu > 0$ there is a noticeable (small) jump in k at the transition. This can be explained if we assume that for $\nu > 0$ the solution scales in a different way with n^{-1}. Close inspection of the jump shows that this jump for $\nu > 0$ is indeed of order $1/\sqrt{n}$, again in perfect agreement with the theory. The difference between the regimes $\nu < 0$ and $\nu > 0$ is also observed in the distribution $\Psi(x)$ of the short-range contributions to the local effective fields; see figure 11.

Although less relevant from a biological point of view, it is interesting to compare the phase diagrams of $n \to \infty$ (or at least large), describing (near-)deterministic sequence selection, to those one would have found for very noisy sequence selection. An example is shown in figure 12, for $n = 1$. Compared to the phase diagram of figure 6, we see that in the presence of high genetic noise the impact of the short-range forces, as measured by J_s is reduced drastically (note the different vertical scales), with as expected a corresponding reduction of sequence selection specificity.

7.2. Numerical simulations

The theory presented in this paper makes a large number of predictions about the cooperative long-time behavior of the polymeric chain. In some asymptotic limits it is possible to work out the expressions for the relevant order parameters of the system and find simplified algebraic

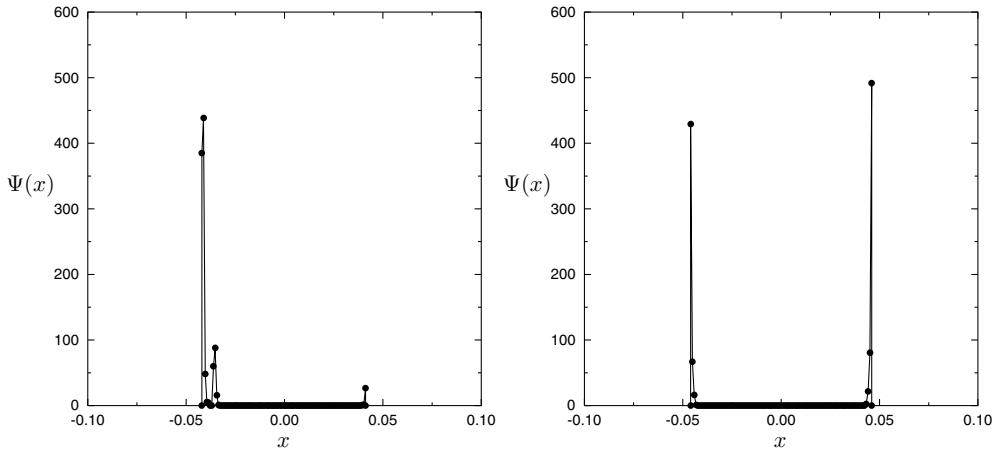

Figure 11. Distribution $\Psi(x)$ of the short-range contributions to the local effective fields, for $(J_s, J_p, J_g) = (0.1, 1, 2)$, $n = 200$, $k_0 = 0.2$, $\mu = 0.7$ and $T = 1.07$, as obtained via a population dynamics algorithm. Left: $\nu = 0.5$. Right: $\nu = -0.5$. Since for $n \to \infty$ the function $\Psi(x)$ is symmetric, so these results confirm that finite-n effects are more profound for $\nu > 0$ (where they are predicted to be $\mathcal{O}(n^{-1/2})$) than for $\nu < 0$ (where they should be $\mathcal{O}(n^{-1})$).

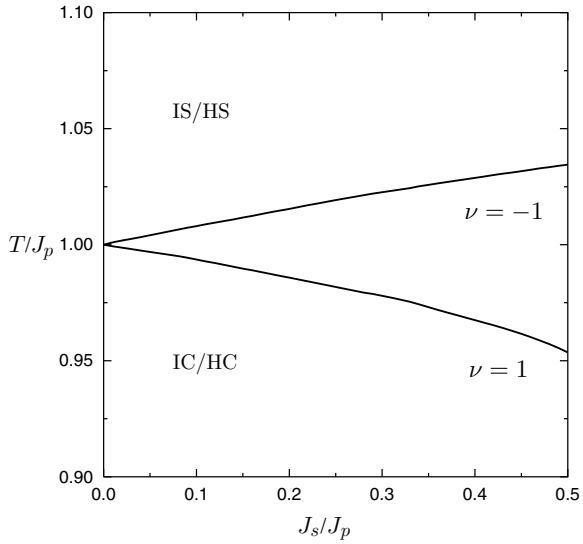

Figure 12. Phase diagram cross-sections for $n = 1$ (strongly noisy sequence selection), $\mu = 0.7$ and $J_g/J_p = 2$. Top curve: $\nu = -1$ (promoting identical orientations of adjacent amino acids). Bottom curve: $\nu = 1$ (promoting opposite orientations). Solid line: transition marking the continuous bifurcation of $m \neq 0$ states. Phases are defined and described in the main text.

equations which allow us to plot phase diagrams. In other cases, we had to rely on population dynamics algorithms to solve our functional order parameter equations and detect the relevant transition lines.

In order to have independent tests of our formulae we have also performed Monte Carlo simulations of the stochastic processes that would lead to equilibration with the Hamiltonians

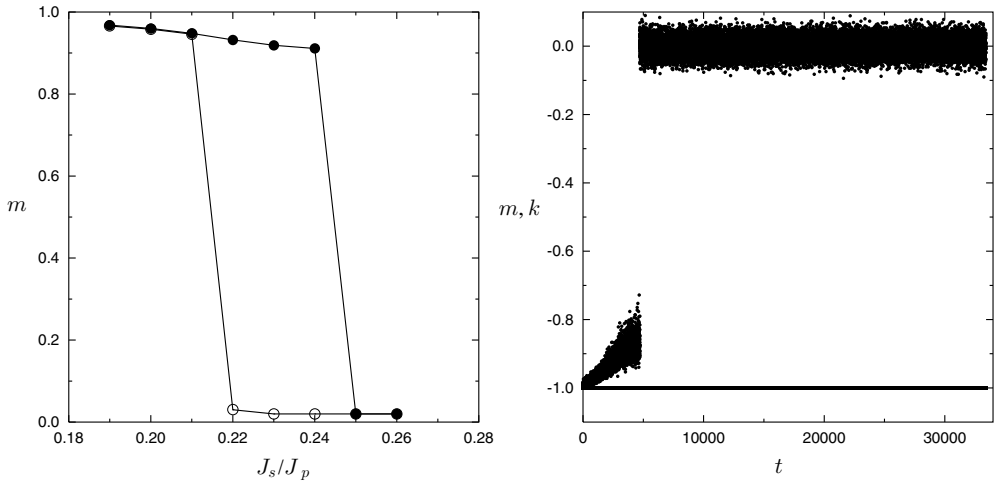

Figure 13. Results of numerical simulations of the coupled stochastic processes of (fast) folding and (slow) primary sequence selection, for $N = 1000$, at $T = 0.3$ and $n = 200$. Further system parameters: $\nu = J_g = \frac{1}{2}$, $J_p = 1$ and $k^\star = 0.7$. For $n \to \infty$ one expects to find the phenomenology of figure 7, with $k = -1$ and with a region where $m = 0$ (swollen) and $m \neq 0$ (folded) states are simultaneously stable; for $n = 200$ one expects this to remain true but with shifted values of J_s/J_p. The left picture shows the equilibrated values of m, found upon increasing J_s in stages from below (full circles) and alternatively upon decreasing J_s in stages from above (open circles). It confirms that there is a coexistence region at the predicted range of values for J_s/J_p. The right picture, measured at $J_s/J_p = 0.25$, shows the evolution in time of m (upper) and k (lower), upon initializing the system in the folded state that is stable for lower values of J_s. It suggests that the chosen duration of 5×10^4 iterations per monomer suffices in the present parameter regime to achieve equilibration.

(1) and (5). This is of course the cleanest way to check the theory. However, due to the special character of the coupled dynamics which requires nested equilibrations of two complex processes at widely separated time scales, these simulations are highly non-trivial and extremely time consuming, and one is severely limited in both the number and the precision of simulation experiments that can be completed reliably. A systematic scan of all possible parameter regimes is certainly ruled out. Instead we focused on the regime $n \to \infty$ (genetic evolution of sequences at low noise levels). This is not only the most relevant one biologically, but is also the regime where our predictions take their most explicit form, as here we could go beyond population dynamics analyses. In particular, we chose to invest our computing resources in verifying the existence and location of the predicted coexistence region in the phase diagram shown in figure 7.

We simulated the coupled Monte Carlo dynamics associated with (1) and (5) for $n = 200$, with $\nu = J_g = \frac{1}{2}$, $J_p = 1$ (so figure 7 is predicted to apply at least in the limit $n \to \infty$) and $k^\star = 0.7$, for a system of $N = 1000$ monomers at folding temperature $T = 0.3$. We employed careful online tests to ensure equilibration of folding angles before carrying out monomer substitutions (i.e. genetic updates), and we allowed for 5×10^4 iterations per monomer. For these parameter choices our $n \to \infty$ theory predicts that always $k = -1$, and that there are two critical values for J_s: one should find $m \neq 0$ (a folded state) for $J_s < 0.181$, $m = 0$ (a swollen state) for $J_s > 0.209$, with coexistence of the $m = 0$ and $m \neq 0$ states for $0.181 < J_s < 0.209$. For large but finite n (here: $n = 200$) one should expect on the basis of the earlier data in, e.g. figure 9 to observe a shift of about 10% in these critical values relative to those of $n \to \infty$. The results of the simulation experiments are shown in figure 13. Each individual point in the

left figure represents the outcome of a simulation where both the fast and the slow processes have equilibrated. The figures confirm the existence of a coexistence region, with critical J_s values compatible with the predicted 10% shift relative to those calculated for $n \to \infty$. The associated value of the order parameter k is indeed $k = -1$. The graph showing the evolution in time of the scalar observables illustrates for $J_s/J_p = 0.25$ how the $m \neq 0$ destabilizes if J_s has become too large, and supports the claim that in the present parameter regime our simulations have equilibrated sufficiently. The remaining fluctuations in m are finite size effects, of the expected order $\Delta m \sim N^{-1/2}$.

8. Discussion

In this paper we have studied the coupled stochastic dynamics of primary and secondary structures formation (i.e. slow-genetic sequence selection and fast folding) in the context of a solvable microscopic model that includes both short-range steric forces and long-range polarity-driven forces. The rationale behind our approach is that it allows us to circumvent the basic obstacle in the application of disordered system techniques to protein folding, which is the need to specify in a mathematical formula the statistics of the disorder, i.e. the statistics of the amino-acid sequences. Here this is not necessary, the sequences are themselves allowed to evolve in time, albeit slowly (to model genetic selection) and in a manner that takes account of the folding properties of the associated chain, and the statistics of sequences are now an implicit *output* of the model rather than an *input*. Our solution is based on exploiting recent mathematical progress [36, 37] in the diagonalization of replicated transfer matrices, and leads in the thermodynamic limit to explicit predictions regarding phase transitions and phase diagrams at genetic equilibrium.

In order to apply the methodology of replicated transfer matrices (which require a formulation in the form of a pseudo-one-dimensional system) we limited ourselves to effective Hamiltonians of a type that represents the physical feasibility and energetic gain of three-dimensional folds indirectly, as in e.g. [26]. Even then, in order to keep the remaining mathematics manageable, we chose to limit ourselves further by retaining only polarity forces and steric forces, we reduced the orientation degrees of freedom of individual monomers, and we made the simplest statistical assumptions regarding polarity and steric properties of amino acids. However, in contrast to the above limitation to pseudo-one-dimensional models, these latter restrictions and choices are not strictly required and can in principle be lifted if one is willing to accept the inevitable associated quantitative increase in mathematical complexity. Even in its reduced form, our model and its solution still have a large number of control parameters to be varied, and a full exploration of its phase phenomenology would have required more than double the present page numbers. Instead we have largely focused on the regime which we believe to be the most relevant one biologically: the large n regime, where the genetic noise is low. We have tried to explain the phases observed and their transitions, and understand these qualitatively.

Our model was found to exhibit a parameter regime where protein-like behavior is observed, i.e. where the genetic selection results in inhomogeneous polarity sequences, and where the folding process describes transitions between swollen and collapsed phases. There was also a parameter regime where the genetic dynamics leads to polymers which are homogeneous in polarity. However, this unbiological behavior requires unphysical values of the control parameters. There is a simple argument to see this. The reason for the energetic advantage of homogeneous polarity sequences is the mean-field contribution $-(J_p/N)\sum_{ij}\xi(\lambda_i)\xi(\lambda_j)\delta_{\phi_i,\phi_j}$ to (1), which even for completely random angles $\{\phi_i\}$, where $\langle\delta_{\phi_i,\phi_j}\rangle = q^{-1}$, retains on average a value $-J_p N(N^{-1}\sum_i \xi(\lambda_i))^2$. In random heteropolymer

models with frozen sequences this term is irrelevant, but here the sequences $\{\lambda_i\}$ evolve, so the system can reduce its energy by increasing $|N^{-1}\sum_i \xi(\lambda_i)|$. A rational alternative definition would be to replace the mean-field term in (1) by $-(J_p/N)\sum_{ij}\xi(\lambda_i)\xi(\lambda_j)[\delta_{\phi_i,\phi_j} - q^{-1}]$, expressing energy gain via folding in terms of *correlation* between side-chain orientations and polarity, rather than *covariance*. This would generate a term similar to the polarity balance energy, and result in the replacement $J_g v(k-k^\star) \to J_g v(k-k^\star) + J_p k^2/q$. For the simple choices $q = 2$, $v(x) = \frac{1}{2}x^2$, and $k^\star = 0$, in particular, the change would translate into the simple parameter re-scaling $J_g \to J_g + J_p$. The natural parameter regime is apparently $J_g > J_p$, that with inhomogeneous polarities.

There is certainly significant scope for improvement and expansion of this study. All our simplifying choices, made for the sake of mathematical convenience, should however be judged in the light of the complexity of the resulting equations even for the presently studied simplified model. The obvious directions to move into next are clear. First, there is the search for more realistic Hamiltonians describing the fast process, by improving the energetic description of the effects of 3D folding (possibly via a formulation involving contact maps, which would replace the long-range all-to-all forces by a sparse connectivity version), and by including hydrogen bonds. Second, we would like to work out our formulae for the case where the monomers' mechanical degrees of freedom consist of two angles, that furthermore can each take more than just two values (preferably a continuum, which would replace the replicated transfer matrices by replicated kernels). Third, one would like to find more realistic alternatives for the sequence selection Hamiltonian, that is more precise in terms of quantifying a sequence's biological functionality, and that employ a better proxy for the unique foldability of a sequence than just its folding free energy.

We see this paper as a proof of principle, demonstrating that it is in principle possible to construct solvable microscopic models of primary and secondary structures formation in heteropolymers, with both long- and short-range forces, in which there is no need to assume (and average over) random amino-acid sequences or to find a formula for suitably non-random sequence statistics. This study represents a small step, but we believe it to be a step in a promising direction.

Acknowledgments

It is our pleasure to thank Isaac Perez-Castillo, Nikos Skantzos and Jort van Mourik for valuable discussions. One of the authors (CJPV) acknowledges financial support from project FIS2006-13321-C02-01 and grant PR2006-0458.

Appendix A. Identification of observables

A.1. 'Slow' free energy as generator of observables

In the stationary state, where both the fast degrees of freedom (ϕ, giving the secondary structure) and the slow degrees of freedom ($\boldsymbol{\lambda}$, giving the primary structure) have equilibrated, expectation values of observables are given by two nested Boltzmann averages. Using definition (5) and $\tilde{\beta} = n\beta$ the result can be written as

$$\langle\langle G(\phi,\boldsymbol{\lambda})\rangle_{\rm fast}\rangle_{\rm slow} = \frac{\sum_{\boldsymbol{\lambda}} {\rm e}^{-\tilde{\beta}H_{\rm eff}(\boldsymbol{\lambda})}\langle G(\phi,\boldsymbol{\lambda})\rangle_{\rm fast}}{\sum_{\boldsymbol{\lambda}} {\rm e}^{-\tilde{\beta}H_{\rm eff}(\boldsymbol{\lambda})}},$$
$$= {\rm e}^{\tilde{\beta}Nf_N}\sum_{\boldsymbol{\lambda}} {\rm e}^{-\tilde{\beta}H_{\rm eff}(\boldsymbol{\lambda})}\left\{\frac{\sum_{\phi}{\rm e}^{-\beta H_{\rm f}(\phi|\boldsymbol{\lambda})}G(\phi,\boldsymbol{\lambda})}{\sum_{\phi}{\rm e}^{-\beta H_{\rm f}(\phi|\boldsymbol{\lambda})}}\right\},$$

$$= \mathrm{e}^{\tilde{\beta} N f_N} \sum_{\lambda} \frac{\mathrm{e}^{-\tilde{\beta}[U(\lambda)+V(\lambda)]}}{\mathcal{Z}_{\mathrm{f}}^{1-n}(\lambda)} \sum_{\phi} G(\phi, \lambda)\, \mathrm{e}^{-\beta H_{\mathrm{f}}(\phi|\lambda)},$$

$$= \mathrm{e}^{\tilde{\beta} N f_N} \sum_{\lambda} \sum_{\phi^1,\ldots,\phi^n} G(\phi^1, \lambda)\, \mathrm{e}^{-\beta \sum_\alpha [H_{\mathrm{f}}(\phi^\alpha|\lambda)+U(\lambda)+V(\lambda)]}, \tag{A.1}$$

with $\alpha = 1, \ldots, n$. This latter expression is also obtained as the derivative of the 'slow' free energy f_N, provided we add a suitable generating term to the 'fast' Hamiltonian $H_{\mathrm{f}}(\phi|\lambda)$. To be precise, upon replacing

$$H_{\mathrm{f}}(\phi|\lambda) \to H_{\mathrm{f}}(\phi|\lambda) + \chi N G(\phi, \lambda), \tag{A.2}$$

one obtains

$$\langle\langle G(\phi, \lambda)\rangle_{\mathrm{fast}}\rangle_{\mathrm{slow}} = \lim_{\chi\to 0} \frac{\partial}{\partial \chi} f_N. \tag{A.3}$$

The validity of (A.3), which allows us to use the free energy as a generating function for expectation values, follows immediately upon substituting (A.2) into (6):

$$\lim_{\chi\to 0} \frac{\partial}{\partial \chi} f_N = -\lim_{\chi\to 0} \frac{1}{nN\beta} \frac{\partial}{\partial \chi} \log \sum_{\lambda} \sum_{\phi^1,\ldots,\phi^n} \mathrm{e}^{-\beta \sum_{\alpha=1}^n [\chi N G(\phi^\alpha,\lambda)+H_{\mathrm{f}}(\phi^\alpha|\lambda)+U(\lambda)+V(\lambda)]}$$

$$= \lim_{\chi\to 0} \frac{1}{n} \sum_{\gamma=1}^{n} \frac{\sum_\lambda \sum_{\phi^1,\ldots,\phi^n} G(\phi^\gamma, \lambda)\, \mathrm{e}^{-\beta \sum_{\alpha=1}^n [\chi G(\phi^\alpha,\lambda)+H_{\mathrm{f}}(\phi^\alpha|\lambda)+U(\lambda)+V(\lambda)]}}{\sum_\lambda \sum_{\phi^1,\ldots,\phi^n} \mathrm{e}^{-\beta \sum_{\alpha=1}^n [\chi G(\phi^\alpha,\lambda)+H_{\mathrm{f}}(\phi^\alpha|\lambda)+U(\lambda+V(\lambda))]}},$$

$$= \frac{\sum_\lambda \sum_{\phi^1,\ldots,\phi^n} G(\phi^1, \lambda)\, \mathrm{e}^{-\beta \sum_{\alpha=1}^n [H_{\mathrm{f}}(\phi^\alpha|\lambda)+U(\lambda)+V(\lambda)]}}{\sum_\lambda \sum_{\phi^1,\ldots,\phi^n} \mathrm{e}^{-\beta \sum_{\alpha=1}^n [H_{\mathrm{f}}(\phi^\alpha|\lambda)+U(\lambda)+V(\lambda)]}},$$

$$= \mathrm{e}^{\tilde{\beta} N f_N} \sum_{\lambda} \sum_{\phi^1,\ldots,\phi^n} G(\phi^1, \lambda)\, \mathrm{e}^{-\beta \sum_{\alpha=1}^n [H_{\mathrm{f}}(\phi^\alpha|\lambda)+U(\lambda)+V(\lambda)]},$$

$$= \langle\langle G(\phi, \lambda)\rangle_{\mathrm{fast}}\rangle_{\mathrm{slow}}. \tag{A.4}$$

A.2. Identification of order parameters for $q = 2$

We next apply the general relations (A.2) and (A.3) for $q = 2$ to observables of the form $G(\boldsymbol{\sigma}, \boldsymbol{\lambda}) = N^{-1} \sum_i g(\sigma_i, \xi_i, \eta_i)$. Here equations (A.2) and (A.3) translate into

$$H_{\mathrm{f}}(\boldsymbol{\sigma}|\boldsymbol{\lambda}) \to H_{\mathrm{f}}(\boldsymbol{\sigma}|\boldsymbol{\lambda}) + \chi \sum_i g(\sigma_i, \xi_i, \eta_i), \tag{A.5}$$

$$\frac{1}{N} \sum_i \langle\langle g(\sigma_i, \xi_i, \eta_i)\rangle_{\mathrm{fast}}\rangle_{\mathrm{slow}} = \lim_{\chi\to 0} \frac{\partial}{\partial \chi} f_N. \tag{A.6}$$

We repeat our previous derivation of the free energy per amino acid (20) but now with the new contribution $\chi \sum_i g(\sigma_i, \xi_i, \eta_i)$ included in the fast Hamiltonian $H_{\mathrm{f}}(\boldsymbol{\sigma})$, in leading order in χ. The new term changes (14) into

$$M[\boldsymbol{\sigma}_{i-1}, \boldsymbol{\sigma}_i, \boldsymbol{\sigma}_{i+1}|\mathbf{m}, \mathbf{k}] = \Big\langle\Big\langle \mathrm{e}^{\beta\xi[J_p \sum_\alpha (k_\alpha + m_\alpha \sigma_i^\alpha) - n\mu]}$$

$$\times\, \mathrm{e}^{-\beta\xi n J_g v'(\frac{1}{n}\sum_\alpha k_\alpha - k^\star) + \beta\eta[J_s \boldsymbol{\sigma}_{i+1}\cdot\boldsymbol{\sigma}_{i-1} - nv] - \beta\chi \sum_\alpha g(\sigma_i^\alpha,\xi,\eta)} \Big\rangle\Big\rangle_{\xi,\eta}. \tag{A.7}$$

From this we can immediately recover the identifications (29) and (30). For instance, choosing $g(\sigma, \xi, \eta) = \sigma\xi$ gives

$$M[\boldsymbol{\sigma}_{i-1}, \boldsymbol{\sigma}_i, \boldsymbol{\sigma}_{i+1}|\mathbf{m}, \mathbf{k}] = M\left[\boldsymbol{\sigma}_{i-1}, \boldsymbol{\sigma}_i, \boldsymbol{\sigma}_{i+1}\middle|\mathbf{m} - \frac{\chi}{J_p}(1, \ldots, 1), \mathbf{k}\right]. \tag{A.8}$$

From this we extract, due to $M[\boldsymbol{\sigma}_{i-1}, \boldsymbol{\sigma}_i, \boldsymbol{\sigma}_{i+1}]$ only affecting the transfer matrix eigenvalue $\lambda(\mathbf{m}, \mathbf{k})$, within the replica symmetric ansatz:

$$\lim_{N\to\infty} \frac{1}{N} \sum_i \langle\langle \xi_i \sigma_i \rangle_{\text{fast}} \rangle_{\text{slow}} = \lim_{\chi\to 0} \frac{\partial}{\partial \chi} f_N = -\frac{1}{\beta n} \lim_{\chi\to 0} \frac{\partial}{\partial \chi} \log \lambda^{\text{RS}}_{\max}\left(m - \frac{\chi}{J_p}, k\right),$$

$$= \frac{1}{\beta n J_p} \frac{\partial}{\partial m} \log \lambda^{\text{RS}}_{\max}(m, k)|_{\chi=0} = m \tag{A.9}$$

according to (22). Similarly, making the alternative choice $g(\sigma, \xi, \eta) = \xi$ gives in leading order in χ:

$$M[\boldsymbol{\sigma}_{i-1}, \boldsymbol{\sigma}_i, \boldsymbol{\sigma}_{i+1}|\mathbf{m}, \mathbf{k}] = M\left[\boldsymbol{\sigma}_{i-1}, \boldsymbol{\sigma}_i, \boldsymbol{\sigma}_{i+1}\middle|\mathbf{m}, \mathbf{k} - \frac{\chi}{J_p}(1, \ldots, 1)\right]_{v'(\cdot)\to v'(\cdot)-\chi v''(\cdot)/J_p}. \tag{A.10}$$

From this we extract, within the replica symmetric ansatz:

$$\lim_{N\to\infty} \frac{1}{N} \sum_i \langle\langle \xi_i \rangle_{\text{fast}} \rangle_{\text{slow}} = \lim_{\chi\to 0} \frac{\partial}{\partial \chi} f_N$$

$$= -\frac{1}{\beta n} \lim_{\chi\to 0} \frac{\partial}{\partial \chi} \log \lambda^{\text{RS}}_{\max}\left(m, k - \frac{\chi}{J_p}\right)\Bigg|_{v'(\cdot)\to v'(\cdot)-\chi v''(\cdot)/J_p},$$

$$= -\frac{1}{\beta n} \lim_{\chi\to 0} \frac{\partial}{\partial \chi} \log \lambda^{\text{RS}}_{\max}\left(m, k - \frac{\chi}{J_p} + \frac{\chi}{J_p}\left[\frac{J_g}{J_p} v''(k - k^\star)\right]\right),$$

$$= \frac{1}{\beta n J_p}\left[1 - \frac{J_g}{J_p} v''(k - k^\star)\right] \frac{\partial}{\partial k} \log \lambda^{\text{RS}}_{\max}(m, k)|_{\chi=0} = k, \tag{A.11}$$

according to (23). The above identification of the scalar order parameters m and k was relatively easy since we could absorb the extra generating terms into those already present. This will generally not be the case.

A.3. Joint distribution of primary structure variables

We next turn to the calculation of the equilibrium amino-acid statistics as measured by $\pi(\hat{\xi}, \hat{\eta}) = \lim_{N\to\infty} N^{-1} \sum_i \langle\langle \delta(\hat{\xi} - \xi_i)\delta(\hat{\eta} - \eta_i)\rangle_{\text{fast}}\rangle_{\text{slow}}$. This distribution follows from (A.5) and (A.6) upon making the choice $g(\sigma, \xi, \eta) = \delta(\xi - \hat{\xi})\delta(\eta - \hat{\eta})$:

$$H_{\text{f}}(\boldsymbol{\sigma}|\boldsymbol{\lambda}) \to H_{\text{f}}(\boldsymbol{\sigma}|\boldsymbol{\lambda}) + \chi \sum_i \delta(\xi_i - \hat{\xi})\delta(\eta_i - \hat{\eta}) \tag{A.12}$$

$$\pi(\hat{\xi}, \hat{\eta}) = \lim_{N\to\infty} \lim_{\chi\to 0} \frac{\partial}{\partial \chi} f_N. \tag{A.13}$$

The calculation is now complicated by the fact that the convenient decomposition identity (15) no longer holds. Instead we now find, in replica symmetric ansatz:

$$M[\boldsymbol{\sigma}_{i-1}, \boldsymbol{\sigma}_i, \boldsymbol{\sigma}_{i+1}|\mathbf{m}, \mathbf{k}] = M[\boldsymbol{\sigma}_{i-1}, \boldsymbol{\sigma}_i, \boldsymbol{\sigma}_{i+1}|\mathbf{m}, \mathbf{k}]|_{\chi=0}$$

$$- n\beta\chi \left\langle \delta(\xi - \hat{\xi})\, \mathrm{e}^{\beta\xi[J_p \sum_\alpha (k + m\sigma_i^\alpha) - n\mu - nJ_g v'(k - k^\star)]} \right\rangle_\xi \left\langle \delta(\eta - \hat{\eta})\, \mathrm{e}^{\beta\eta[J_s \boldsymbol{\sigma}_{i+1}\cdot\boldsymbol{\sigma}_{i-1} - n\nu]} \right\rangle_\eta$$

$$+ \mathcal{O}(\chi^2), \tag{A.14}$$

so that

$$\prod_i M[\boldsymbol{\sigma}_{i-1},\boldsymbol{\sigma}_i,\boldsymbol{\sigma}_{i+1}|\mathbf{m},\mathbf{k}] = \prod_i \Gamma_{\boldsymbol{\sigma}_{i-1},\boldsymbol{\sigma}_{i+1}}(\mathbf{m},\mathbf{k}) + \mathcal{O}(\chi^2)$$
$$- n\beta\chi \sum_j \left\{ \left\langle \delta(\xi-\hat{\xi})\, \mathrm{e}^{\beta\xi[J_p\sum_\alpha (k+m\sigma_i^\alpha)-n\mu-nJ_g v'(k-k^\star)]}\right\rangle_\xi \right.$$
$$\left. \times \left\langle \delta(\eta-\hat{\eta})\, \mathrm{e}^{\beta\eta[J_s\boldsymbol{\sigma}_{j+1}\cdot\boldsymbol{\sigma}_{j-1}-n\nu]}\right\rangle_\eta \prod_{i\neq j} M[\boldsymbol{\sigma}_{i-1},\boldsymbol{\sigma}_i,\boldsymbol{\sigma}_{i+1}|\mathbf{m},\mathbf{k}] \right\}$$
$$= \left[\prod_i \Gamma_{\boldsymbol{\sigma}_{i-1},\boldsymbol{\sigma}_{i+1}}(\mathbf{m},\mathbf{k})\right]$$
$$\times \left[1 - n\beta\chi \sum_j V_{\boldsymbol{\sigma}_j}(m,k) W_{\boldsymbol{\sigma}_{j-1}\boldsymbol{\sigma}_{j+1}}(m,k) + \mathcal{O}(\chi^2)\right], \tag{A.15}$$

with

$$V_{\boldsymbol{\sigma}}(m,k) = \frac{\left\langle \delta(\xi-\hat{\xi})\, \mathrm{e}^{n\beta\xi[\frac{J_p}{n}\sum_\alpha (k+m\sigma_i^\alpha)-\mu-J_g v'(k-k^\star)]}\right\rangle_\xi}{\left\langle \mathrm{e}^{n\beta\xi[\frac{J_p}{n}\sum_\alpha (k+m\sigma_i^\alpha)-\mu-J_g v'(k-k^\star)]}\right\rangle_\xi}, \tag{A.16}$$

$$W_{\boldsymbol{\sigma}\boldsymbol{\sigma}'}(m,k) = \frac{\left\langle \delta(\eta-\hat{\eta})\, \mathrm{e}^{n\beta\eta[\frac{J_s}{n}\boldsymbol{\sigma}\cdot\boldsymbol{\sigma}'-\nu]}\right\rangle_\eta}{\left\langle \mathrm{e}^{n\beta\eta[\frac{J_s}{n}\boldsymbol{\sigma}\cdot\boldsymbol{\sigma}'-\nu]}\right\rangle_\eta}. \tag{A.17}$$

This leads us to

$$\lim_{N\to\infty}\frac{1}{N}\log \sum_{\boldsymbol{\sigma}_1,\ldots,\boldsymbol{\sigma}_N}\prod_i M[\cdots|\cdots] = \log\lambda_{\max}^{\mathrm{RS}}(m,k)|_{\chi=0} \tag{A.18}$$

$$-n\beta\chi \lim_{N\to\infty}\frac{1}{N}\sum_j \frac{\sum_{\boldsymbol{\sigma}_1,\ldots,\boldsymbol{\sigma}_N}\left[\prod_i \Gamma_{\boldsymbol{\sigma}_{i-1}\boldsymbol{\sigma}_{i+1}}(\mathbf{m},\mathbf{k})\right] V_{\boldsymbol{\sigma}_j}(m,k) W_{\boldsymbol{\sigma}_{j-1}\boldsymbol{\sigma}_{j+1}}(m,k)}{\sum_{\boldsymbol{\sigma}_1\ldots\boldsymbol{\sigma}_N}\left[\prod_i \Gamma_{\boldsymbol{\sigma}_{i-1}\boldsymbol{\sigma}_{i+1}}(\mathbf{m},\mathbf{k})\right]} + \mathcal{O}(\chi^2) \tag{A.19}$$

and hence, with $\boldsymbol{\Gamma}(m,k)$ denoting the replica symmetric version (21) of the transfer matrix $\boldsymbol{\Gamma}(\mathbf{m},\mathbf{k})$, with $\lambda_{\max}^{\mathrm{RS}}(m,k)$ denoting the largest eigenvalue of $\boldsymbol{\Gamma}(m,k)$, and using the periodicity of the chain:

$$\lim_{N\to\infty} f_N = \mathrm{extr}_{m,k}\left\{\frac{1}{2}J_p(m^2+k^2) + J_g[v(k-k^\star) - kv'(k-k^\star)]\right.$$
$$+ \chi \lim_{N\to\infty}\frac{\sum_{\boldsymbol{\sigma}_1\ldots\boldsymbol{\sigma}_N}\left[\prod_i \Gamma_{\boldsymbol{\sigma}_{i-1}\boldsymbol{\sigma}_{i+1}}(\mathbf{m},\mathbf{k})\right] V_{\boldsymbol{\sigma}_1}(m,k) W_{\boldsymbol{\sigma}_N\boldsymbol{\sigma}_2}(m,k)}{(\mathrm{Tr}[\boldsymbol{\Gamma}^{N/2}(m,k)])^2}$$
$$\left. - \frac{1}{\beta n}\log\Lambda - \frac{1}{\beta n}\log\lambda_{\max}^{\mathrm{RS}}(m,k) + \mathcal{O}(\chi^2)\right\}, \tag{A.20}$$

$$\pi(\hat{\xi},\hat{\eta}) = \left[\lim_{N\to\infty}\frac{\sum_{\boldsymbol{\sigma}_1\boldsymbol{\sigma}_3}\boldsymbol{\Gamma}^{N/2-1}_{\boldsymbol{\sigma}_3\boldsymbol{\sigma}_1}(m,k)\Gamma_{\boldsymbol{\sigma}_1\boldsymbol{\sigma}_3}(m,k)V_{\boldsymbol{\sigma}_1}(m,k)}{\mathrm{Tr}[\boldsymbol{\Gamma}^{N/2}(m,k)]}\right]$$
$$\times \left[\lim_{N\to\infty}\frac{\sum_{\boldsymbol{\sigma}_2\boldsymbol{\sigma}_N}\boldsymbol{\Gamma}^{N/2-1}_{\boldsymbol{\sigma}_2\boldsymbol{\sigma}_N}(m,k)\Gamma_{\boldsymbol{\sigma}_N\boldsymbol{\sigma}_2}(m,k)W_{\boldsymbol{\sigma}_N\boldsymbol{\sigma}_2}(m,k)}{\mathrm{Tr}[\boldsymbol{\Gamma}^{N/2}(m,k)]}\right]. \tag{A.21}$$

In the latter expression one must substitute for (m, k) the solution of the original $\chi = 0$ saddle-point problem. We find once more a convenient effective decoupling of the odd sites from the even sites, as well as statistical independence of the single-site polarity and steric angle statistics, giving $\pi(\hat{\xi}, \hat{\eta}) = \pi(\hat{\xi})\pi(\hat{\eta})$ with the individual distributions

$$\pi(\hat{\xi}) = \lim_{N\to\infty} \frac{\sum_{\sigma\sigma'} \mathbf{\Gamma}^{N/2-1}_{\sigma'\sigma}(m,k)\left\langle \delta(\xi-\hat{\xi})\, \mathrm{e}^{\beta\xi[J_p(nk+m\sum_\alpha \sigma_i^\alpha)-n\mu-nJ_g v'(k-k^\star)]}\right\rangle_\xi \left\langle \mathrm{e}^{\beta\eta[J_s\boldsymbol{\sigma}\cdot\boldsymbol{\sigma}'-n\nu]}\right\rangle_\eta}{\mathrm{Tr}[\mathbf{\Gamma}^{N/2}(m,k)]}, \tag{A.22}$$

$$\pi(\hat{\eta}) = \lim_{N\to\infty} \frac{\sum_{\sigma\sigma'} \mathbf{\Gamma}^{N/2-1}_{\sigma'\sigma}(m,k)\left\langle \mathrm{e}^{\beta\xi[J_p(nk+m\sum_\alpha \sigma_i^\alpha)-n\mu-nJ_g v'(k-k^\star)]}\right\rangle_\xi \left\langle \delta(\eta-\hat{\eta})\, \mathrm{e}^{\beta\eta[J_s\boldsymbol{\sigma}\cdot\boldsymbol{\sigma}'-n\nu]}\right\rangle_\eta}{\mathrm{Tr}[\mathbf{\Gamma}^{N/2}(m,k)]}. \tag{A.23}$$

The limit $N \to \infty$ can now be taken upon using the fact that for $N \to \infty$ one may write in leading order $\mathbf{\Gamma}^N_{\sigma\sigma'}(m,k) \to \lambda^N(m,k) u^{\mathrm{R}}_\sigma u^{\mathrm{L}}_{\sigma'} \big/ \sum_{\sigma''} u^{\mathrm{L}}_{\sigma''} u^{\mathrm{R}}_{\sigma''}$, where $\{u^{\mathrm{L}}_\sigma\}$ and $\{u^{\mathrm{R}}_\sigma\}$ denote the left and right eigenvectors of $\mathbf{\Gamma}(m,k)$ associated with the largest eigenvalue. In the result we can then substitute our expression (41) for the largest eigenvalue and the replica symmetric forms (31) and (32) for the eigenvectors. For the polarity distribution $\pi(\hat{\xi})$ this gives, after some further manipulations and with help of the definitions (45) and (52):

$$\begin{aligned}
\pi(\hat{\xi}) &= \frac{\sum_{\sigma\sigma'} u^{\mathrm{L}}_\sigma u^{\mathrm{R}}_{\sigma'} \left\langle \delta(\xi-\hat{\xi})\, \mathrm{e}^{\beta\xi[J_p(nk+m\sum_\alpha \sigma_i^\alpha)-n\mu-nJ_g v'(k-k^\star)]}\right\rangle_\xi \left\langle \mathrm{e}^{\beta\eta[J_s\boldsymbol{\sigma}\cdot\boldsymbol{\sigma}'-n\nu]}\right\rangle_\eta}{\lambda(m,k)\sum_\sigma u^{\mathrm{L}}_\sigma u^{\mathrm{R}}_\sigma}, \\
&= \frac{p(\hat{\xi})\int \mathrm{d}h \cosh^n(\beta h)\int \mathrm{d}x\, \Psi(x)\Psi(h-x-\hat{\xi}J_p m)\rangle_\xi}{\int \mathrm{d}\xi\; p(\xi)\int \mathrm{d}h \cosh^n(\beta h)\int \mathrm{d}x\, \Psi(h-x-\xi J_p m)\Psi(x)}, \\
&= \int \mathrm{d}h\, W(\hat{\xi}, h).
\end{aligned} \tag{A.24}$$

For the steric angle distribution $\pi(\hat{\eta})$ one finds an expression with a similar structure:

$$\begin{aligned}
\pi(\hat{\eta}) &= \frac{\sum_{\sigma\sigma'} u^{\mathrm{L}}_\sigma u^{\mathrm{R}}_{\sigma'} \left\langle \mathrm{e}^{\beta\xi[J_p(nk+m\sum_\alpha \sigma_\alpha)-n\mu-nJ_g v'(k-k^\star)]}\right\rangle_\xi \left\langle \delta(\eta-\hat{\eta})\, \mathrm{e}^{\beta\eta[J_s\boldsymbol{\sigma}\cdot\boldsymbol{\sigma}'-n\nu]}\right\rangle_\eta}{\lambda(m,k)\sum_\sigma u^{\mathrm{L}}_\sigma u^{\mathrm{R}}_\sigma}, \\
&= \frac{\int \mathrm{d}x\, \mathrm{d}x'\, \Phi(x')\Psi(x)\langle\!\langle \delta(\eta-\hat{\eta})\, \mathrm{e}^{n\beta[B(x',\eta J_s)+\xi(J_p k-\mu-J_g v'(k-k^\star))-\eta\nu]} \cosh^n[\beta(x+\xi J_p m+A(x',\eta J_s))]\rangle\!\rangle_{\xi,\eta}}{\int \mathrm{d}x\, \Phi(x)\langle\!\langle \mathrm{e}^{n\beta[B(x,\eta J_s)+\xi(J_p k-\mu-J_g v'(k-k^\star))-\nu\eta]}\rangle\!\rangle_{\xi,\eta}\int \mathrm{d}x\, \mathrm{d}x'\, \Phi(x')\Psi(x)\cosh^n[\beta(x+x')]}, \\
\pi(\hat{\eta}) &= \frac{\int \mathrm{d}h \cosh^n(\beta h)\int \mathrm{d}x\, \Phi(x)\Phi(h-A(x,\hat{\eta}J_s))\langle \delta(\eta-\hat{\eta})\, \mathrm{e}^{n\beta[B(x,\eta J_s)-\eta\nu]}\rangle_\eta}{\int \mathrm{d}h \cosh^n(\beta h)\int \mathrm{d}x\, \Phi(x)\Phi(h-A(x,\hat{\eta}J_s))\left\langle \mathrm{e}^{n\beta[B(x,\eta J_s)-\eta\nu]}\right\rangle_\eta}.
\end{aligned} \tag{A.25}$$

Both in the limit $n \to 0$ (fully random sequence selection) and in the limit $\beta \to 0$ one sees both equilibrated distributions reducing to the prior statistics $w(\hat{\xi})$ and $w(\hat{\eta})$, as it should. In general, however, one will find non-trivial distributions $\pi(\hat{\xi})$ and $\pi(\hat{\eta})$, which reflect the complicated interplay between the secondary and primary structures generation. Finally we observe that $\pi(\xi) \neq p(\xi)$, except when $J_p m = 0$; this suggests that, rather than the polarity distribution in the equilibrated system, the physical interpretation of $p(\xi)$ is that of a prior distribution which would have been found in the absence of the secondary structure formation.

Appendix B. Saddle-point treatment of order parameter equations in the limit $n \to \infty$

B.1. Saddle-point treatment of the equation for $\Psi(x)$

Since we know that $\Psi(x) = 0$ for $|x| > J_s$, we may write without loss of generality $\Psi(x) = \mathrm{e}^{n\beta\psi(x)}$ for $x \in \Omega \subseteq [-J_s, J_s]$ and $\Psi(x) = 0$ for $x \notin \Omega$, where $\int_\Omega \mathrm{d}x \mathrm{e}^{n\beta\psi(x)} = 1$. We

also define for $|x| \leqslant y$ and $y > 0$ the function

$$C(x, y) = \frac{1}{\beta} \tanh^{-1}[\tanh(\beta x)/\tanh(\beta y)]. \tag{B.1}$$

It is the inverse of the function $A(x, y)$ with respect to the variable x, since $C(A(x, y), y) = x$ for all $|x| < |y|$. We note that $C(0, y) = 0$ and $\mathrm{sgn}[C(x, y)] = \mathrm{sgn}(x)$. We can now insert our expression for $p(\xi)$ and the definition $\Psi(x) = \mathrm{e}^{n\beta\psi(x)}$ (for $x \in \Omega$, with $\Psi(x) = 0$ elsewhere) into our equation for $\Psi(x)$, and use the function $C(x, y)$ to subsequently transform variables inside the δ-distribution on the right-hand side. Since the Jacobian of this transformation will not be exponential in n as $n \to \infty$, as a result of these manipulations we find for all $x \in \Omega$ an equation for $\psi(x)$ that is for $n \to \infty$ evaluated by the steepest descent:

$$\lim_{n\to\infty} \psi(x) = \lim_{n\to\infty} \frac{1}{\beta n} \log \left\{ \frac{\int_\Omega \mathrm{d}y \int \mathrm{d}\xi\, \mathrm{d}\eta\, w(\eta) w(\xi) \delta\big[C(x, \eta J_s) - y - J_p m\xi\big] \mathrm{e}^{n\beta[\psi(y)+\xi(J_p-J_g)(k-k_0)+B(y+J_p m\xi, \eta J_s)-\nu\eta]}}{\int_\Omega \mathrm{d}y \int \mathrm{d}\xi\, \mathrm{d}\eta w(\eta) w(\xi)\, \mathrm{e}^{n\beta[\psi(y)+\xi(J_p-J_g)(k-k_0)+B(y+J_p m\xi, \eta J_s)-\nu\eta]}} \right\} \tag{B.2}$$

$$= \max_{y\in\Omega, y=C(x,\eta J_s)-J_p m\xi, |\xi|\leqslant 1, |\eta|\leqslant 1} \{\psi(y) + \xi(J_p - J_g)(k - k_0) + B(y + J_p m\xi, \eta J_s) - \nu\eta\}$$
$$- \max_{y\in\Omega, |\xi|\leqslant 1, |\eta|\leqslant 1} \{\psi(y) + \xi(J_p - J_g)(k - k_0) + B(y + J_p m\xi, \eta J_s) - \nu\eta\}. \tag{B.3}$$

Solving the optimization problem (B.3) means calculating both the set $\Omega \subseteq [-J_s, J_s]$ and the function $\lim_{n\to\infty} \psi(x)$ for $x \in \Omega$. Let us inspect some properties of this optimization problem in more detail. Since the maximization in the first line of (B.3) is over a *subset* of the set in the second line (instead of allowing for all $y \in \Omega$, in the first line we impose $y = C(x, \eta J_s) - J_p m\xi$), it is inevitable that $\lim_{n\to\infty} \psi(x) \leqslant 0$ for all $x \in \Omega$. We now know that $\psi_{\max} = \lim_{n\to\infty} \max_{x\in\Omega} \psi(x) \leqslant 0$. This leaves two options: $\psi_{\max} < 0$ versus $\psi_{\max} = 0$. In the first case, however, we would get $\lim_{n\to\infty} \Psi(x) = \lim_{n\to\infty} \mathrm{e}^{n\beta\psi(x)} \leqslant \lim_{n\to\infty} \mathrm{e}^{n\beta\psi_{\max}} = 0$ for all $x \in \Omega$; this function can never be normalized. We conclude that $\psi_{\max} = 0$.

Let us turn to those values of x for which one has $\lim_{n\to\infty} \psi(x) = \psi_{\max} = 0$. We call the set of those values $\Omega^\star \subseteq \Omega$:

$$x \in \Omega^\star: \max_{y\in\Omega, |\xi|, |\eta|\leqslant 1, x=A(y+J_p m\xi, \eta J_s)} \{\psi(y) + \xi(J_p - J_g)(k - k_0) + B(y + J_p m\xi, \eta J_s) - \nu\eta\}$$
$$= \max_{y\in\Omega, |\xi|, |\eta|\leqslant 1} \{\psi(y) + \xi(J_p - J_g)(k - k_0) + B(y + J_p m\xi, \eta J_s) - \nu\eta\}. \tag{B.4}$$

We see that with every combination (y, ξ, η) that gives the maximum value in the second line there corresponds a value of $x \in \Omega^\star$. If the maximum is obtained for a *unique* combination $(y^\star, \xi^\star, \eta^\star)$, which apart from symmetries one must expect to be the generic case, then the set $\Omega^\star$ contains just one element $x^\star = A(y^\star + J_p m\xi^\star, \eta^\star J_s)$. It follows that one must generally anticipate $\lim_{n\to\infty} \Psi(x)$ to be a sum of a small number of δ-peaks.

We can finally also use saddle-point arguments to express the limit $n \to \infty$ of the free energy per monomer (54) in terms of the function $\psi(x)$, the scalar order parameters (k, m), and the set Ω:

$$\lim_{n\to\infty} \varphi = \tfrac{1}{2} J_p(m^2 + k^2) - \tfrac{1}{2} J_g(k^2 - k^{\star 2}) - |J_p - J_g||k - k_0|$$
$$- \max_{x\in\Omega, \xi, \eta\in[-1,1]} \{\psi(x) + \xi(J_p - J_g)(k - k_0) + B(x + J_p m\xi, \eta J_s) - \nu\eta\}. \tag{B.5}$$

In the remainder of this section we will not attempt to solve problem (B.3) in its full generality, but rather construct two qualitatively different specific solutions of (B.3), for which

indeed $\Psi(x)$ is found to reduce to either one or two δ-peaks, and which both reduce exactly to the unique solutions that we established earlier in the two limits $J_s \to 0$ or $J_p m \to 0$.

B.2. Homogeneous polarity states $k = \pm 1$

Here we construct solutions of (B.3) where $\Omega = \{x^\star\}$, and show that these represent the continuation to arbitrary $J_s > 0$ and $J_p m \neq 0$ of the homogeneous polarity states $k = \pm 1$. Now we must have $\Omega^\star = \Omega$ and $\lim_{n\to\infty} \psi(x^\star) = \psi_{\max} = 0$, and (B.4) becomes

$$\max_{\xi,\eta\in[-1,1],x^\star=A(x^\star+J_p m\xi,\eta J_s)} L(\xi,\eta) = \max_{\xi,\eta\in[-1,1]} L(\xi,\eta), \tag{B.6}$$

$$L(\xi,\eta) = \xi(J_p - J_g)(k-k_0) + B(x^\star + J_p m\xi, \eta J_s) - \nu\eta. \tag{B.7}$$

In both sides of (B.6) we maximize exactly the same object, but on the left-hand side we have the additional constraint that the values (ξ,η) for which the maximum is found *must* allow the equation $x^\star = A(x^\star + J_p m\xi, \eta J_s)$ to have a solution $x^\star \in [-J_s, J_s]$. If the maximum on the (less constrained) right-hand side is obtained for a (ξ,η) such that the equation $x^\star = A(x^\star + J_p m\xi, \eta J_s)$ has *no* solution $x^\star \in [-J_s, J_s]$, then the extra constraint apparently interferes with the maximization and the two sides *cannot* be the same, so no solution with $\Omega = \{x^\star\}$ can exist. We conclude that the present type of solution exists if and only if both sides of (B.6) find their maximum at the same value $(\hat\xi, \hat\eta)$ (values that will depend on $x^\star$, since $x^\star$ appears in the function to be maximized), with the value of $x^\star$ subsequently following from solution of the nonlinear equation $x^\star = A(x^\star + J_p m\hat\xi, \hat\eta J_s)$:

$$(\hat\xi, \hat\eta) = \text{argmax}_{\xi,\eta\in[-1,1]} L(\xi,\eta), \tag{B.8}$$

$$x^\star = A(x^\star + J_p m\hat\xi, \hat\eta J_s). \tag{B.9}$$

Since $B(x,y) = B(|x|,|y|)$, and is monotonically increasing with both $|x|$ and $|y|$ we can immediately maximize with respect to $\eta \in [-1,1]$, giving $\hat\eta = -\text{sgn}(\nu)$. This simplifies our remaining problem to solving

$$x^\star = -\text{sgn}(\nu) A(x^\star + J_p m\hat\xi, J_s), \tag{B.10}$$

$$\hat\xi = \text{argmax}_{\xi\in[-1,1]} L(\xi), \tag{B.11}$$

$$L(\xi) = \xi(J_p - J_g)(k-k_0) + B(x^\star + J_p m\xi, J_s). \tag{B.12}$$

To resolve the remaining extremization we inspect the properties of $B(x,y)$, in particular its second partial derivative in x. We find that the function $L(\xi)$ is convex:

$$\frac{\partial^2 L(\xi)}{\partial \xi^2} = J_p^2 m^2 \left\{ 1 - \frac{1}{2}\tanh^2[\beta(x^\star + J_p m\xi + J_s)] - \frac{1}{2}\tanh^2[\beta(x^\star + J_p m\xi - J_s)] \right\} \geqslant 0. \tag{B.13}$$

$L(\xi)$ can therefore only be maximal at the boundaries $\xi \in \{-1,1\}$. Next we can rule out states with $x^\star = 0$, since substitution into (B.10) shows that they would be incompatible with $\hat\xi = \pm 1$. Due to $J_p m \neq 0$, $x^\star \neq 0$, and the monotonicity and symmetry of $B(x,y)$, the function $L(\xi)$ is not symmetric in ξ, hence its maximum is unique:

$$\hat\xi = \text{sgn}\left\{ (J_p - J_g)(k-k_0) + \frac{1}{2}[B(x^\star + J_p m, J_s) - B(x^\star - J_p m, J_s)] \right\}. \tag{B.14}$$

Having solved the extremization problem for solutions with $\Omega = \{x^\star\}$, resulting in the two coupled equations (B.10) and (B.14) we turn to the $n \to \infty$ limit of the order parameter equations (81) and (82) for m and k. We define

$$R(\xi) = \xi(J_p - J_g)(k - k_0) + \frac{1}{\beta}\log\cosh[\beta(J_p m\xi + 2x^\star)]. \tag{B.15}$$

This is again a convex function, which is asymmetric in ξ (due to $x^\star \neq 0$), and therefore takes its maximal value on the interval $[-1, 1]$ at the boundary

$$\bar{\xi} = \mathrm{sgn}\left\{(J_p - J_g)(k - k_0) + \frac{1}{2\beta}\log\left[\frac{\cosh[\beta(J_p m + 2x^\star)]}{\cosh[\beta(J_p m - 2x^\star)]}\right]\right\}. \tag{B.16}$$

Our equations for m and k can now be written as

$$m = \lim_{n\to\infty}\frac{\int \mathrm{d}\xi\; w(\xi)\xi\tanh[\beta(J_p m\xi + 2x^\star)]\mathrm{e}^{n\beta R(\xi)}}{\int \mathrm{d}\xi\; w(\xi)\,\mathrm{e}^{n\beta R(\xi)}},$$

$$= \tanh[\beta(J_p m + 2x^\star\bar{\xi})], \tag{B.17}$$

$$k = \lim_{n\to\infty}\frac{\int \mathrm{d}\xi\; w(\xi)\xi\,\mathrm{e}^{n\beta R(\xi)}}{\int \mathrm{d}\xi\; w(\xi)\,\mathrm{e}^{n\beta R(\xi)}} = \bar{\xi}. \tag{B.18}$$

We have now confirmed that the present family of solutions with $\Omega = \{x^\star\}$ are indeed the generalization to arbitrary values of J_s and $J_p m$ of the solutions $k = \pm 1$ with homogeneous polarity, as claimed. Putting all our final equations together, replacing $\bar{\xi}$ by $k \in \{-1, 1\}$ and using the full definition of $B(x, y)$, gives the new set

$$x^\star = -\mathrm{sgn}(v)A(x^\star + J_p m\hat{\xi}, J_s), \tag{B.19}$$

$$m = \tanh[\beta(J_p m + 2x^\star k)], \tag{B.20}$$

$$k = \mathrm{sgn}\left\{(J_p - J_g)(k - k_0) + \frac{1}{2\beta}\log\left[\frac{\cosh[\beta(2x^\star + J_p m)]}{\cosh[\beta(2x^\star - J_p m)]}\right]\right\}, \tag{B.21}$$

$$\hat{\xi} = \mathrm{sgn}\left\{(J_p - J_g)(k - k_0) + \frac{1}{4\beta}\log\left[\frac{\cosh[\beta(x^\star + J_p m + J_s)]\cosh[\beta(x^\star + J_p m - J_s)]}{\cosh[\beta(x^\star - J_p m + J_s)]\cosh[\beta(x^\star - J_p m - J_s)]}\right]\right\}. \tag{B.22}$$

In both of the limits $J_s \to 0$ and $J_p m \to 0$ we recover correctly the equations of the $k = \pm 1$ states as derived earlier for these special cases, namely $x^\star = 0, m = \tanh(\beta J_p m)$, and $k = \hat{\xi} = \mathrm{sgn}[(J_p - J_g)(k - k_0)]$.

Finally we try to compactify and simplify our equations. We first solve $x^\star$ from (B.20), which gives

$$x^\star = k\left[\frac{1}{2\beta}\tanh^{-1}(m) - \frac{1}{2}J_p m\right]. \tag{B.23}$$

Subsequent insertion into (B.19) leaves us with

$$\frac{\tanh\left[\frac{1}{2}\mathrm{arctanh}(m) - \frac{1}{2}\beta J_p m\right]}{\tanh\left[\frac{1}{2}\mathrm{arctanh}(m) - \frac{1}{2}\beta J_p m(1 - 2k\hat{\xi})\right]} = -\mathrm{sgn}(v)\tan(\beta J_s). \tag{B.24}$$

Furthermore, we note that with $k\hat{\xi} \in \{-1, 1\}$ only the choice $k = \hat{\xi}$ will allow the above equations to reduce to the equations for $m \to 0$ that were found earlier, and that the alternative

$k = -\hat{\xi}$ would make it extremely difficult to satisfy both (B.21) and (B.22) simultaneously. Upon choosing $\hat{\xi} = k$ and after additional rearranging and manipulation we can reduce our set of equations further to

$$\mathrm{sgn}(\nu)\tan(\beta J_s) = \frac{\tanh\left[\frac{1}{2}\beta J_p|m| - \frac{1}{2}\tanh^{-1}|m|\right]}{\tanh\left[\frac{1}{2}\beta J_p|m| + \frac{1}{2}\tanh^{-1}|m|\right]}, \tag{B.25}$$

$$(J_p - J_g)(1 - k_0 k) > \frac{1}{2\beta}\log\left[\frac{\cosh[\tanh^{-1}|m| - 2\beta J_p|m|]}{\cosh[\mathrm{arctanh}|m|]}\right], \tag{B.26}$$

$$(J_p - J_g)(1 - k_0 k) > \frac{1}{4\beta}\log\left[\frac{\cosh[\mathrm{arctanh}|m| - 3\beta J_p|m|] + \cosh(2\beta J_s)}{\cosh[\mathrm{arctanh}|m| + \beta J_p|m|] + \cosh(2\beta J_s)}\right]. \tag{B.27}$$

The joint distribution $W(h, \xi)$ of effective fields and polarities for the present solution is very simple:

$$W(h, \xi) = \delta[h - k\beta^{-1}\tanh^{-1}(m)]\delta(\xi - k) \tag{B.28}$$

Working out the free energy per monomer (B.5) for the above solution gives, using $k_0 \in (-1, 1)$ and equation (B.23) to eliminate $x^\star$:

$$\lim_{n\to\infty}\varphi = \frac{1}{2}J_p(m^2+1) - \frac{1}{2}J_g(1-k^{\star 2}) - |J_p - J_g|(1 - kk_0) - (J_p - J_g)(1 - kk_0)$$
$$- |\nu| - B\left(\frac{1}{2\beta}\tanh^{-1}(m) + \frac{1}{2}J_p m, J_s\right) \tag{B.29}$$

Equation (B.25) gives a single transparent law from which to solve our order parameter m. Equations (B.20) and (B.21) give conditions for the solution of (B.25) to be acceptable; they are guaranteed to be satisfied for small m if $J_p > J_g$ (due to $|k_0| < 1$), whereas for larger m their validity needs to be checked explicitly. Equations (B.20) and (B.21) also suggest that, as was found explicitly in the simple cases $J_s = 0$ and $J_p m = 0$, the most stable solution (and hence the thermodynamic state) will generally be the one with $k = -\mathrm{sgn}(k_0)$. This completes our analysis of solutions with $\Omega = \{x^\star\}$. We always find $k = \pm 1$, namely sequences with homogeneous polarity, provided $J_p > J_g$.

B.3. Inhomogeneous polarity states $k = k_0$

In the same manner we now construct the continuation to arbitrary values of J_s and $J_p m$ of the inhomogeneous polarity states, where $k = k_0$. For this case, where Ω no longer contains just one point, our equation (B.3) from which to solve $\lim_{n\to\infty}\psi(x)$ takes the following form:

$$\psi(x) = \max_{\xi,\eta\in[-1,1], y\in\Omega, y=C(x,\eta J_s)-J_p m\xi} L(\xi, \eta, y) - \max_{\xi,\eta\in[-1,1], y\in\Omega} L(\xi, \eta, y), \tag{B.30}$$

$$L(\xi, \eta, y) = \psi(y) + B(y + J_p m\xi, \eta J_s) - \nu\eta \tag{B.31}$$

(provided $x \in \Omega$). In contrast to the $k \neq k_0$ case, this equation has symmetries that can be exploited: it allows for solutions with $\psi(-x) = \psi(x)$ for all $x \in \Omega$, with Ω being symmetric around the origin. This is easily confirmed by working out the right-hand side of (B.30) under the assumption of this symmetry (via transformations like $y \to -y$ and $\xi \to -\xi$, which are allowed by the constraints) upon making the replacement $x \to -x$ on the left-hand side and

using $L(-\xi, \eta, -y) = L(\xi, \eta, y)$ and $C(-x, y) = C(x, y)$:

$$\begin{aligned}\psi(-x) - \psi(x) &= \max_{\xi,\eta\in[-1,1], y\in\Omega, y=C(-x,\eta J_s)-J_p m\xi} L(\xi, \eta, y) \\ &\quad - \max_{\xi,\eta\in[-1,1], y\in\Omega, y=C(x,\eta J_s)-J_p m\xi} L(\xi, \eta, y), \\ &= \max_{\xi,\eta\in[-1,1], y\in\Omega, y=C(x,\eta J_s)-J_p m\xi} L(-\xi, \eta, -y) \\ &\quad - \max_{\xi,\eta\in[-1,1], y\in\Omega, y=C(x,\eta J_s)-J_p m\xi} L(\xi, \eta, y) = 0. \end{aligned} \tag{B.32}$$

We will now construct solutions for $k = k_0$ with this reflection symmetry. Inside (B.30) it allows us to transform without punishment $y \to y\,\mathrm{sgn}(\eta)$ and $\xi \to \xi\,\mathrm{sgn}(\eta)$, which gives a new expression that shows (using $C(x, -y) = -C(x, y)$ and $B(x, y) = B(|x|, |y|)$) that both terms are maximized for $\mathrm{sgn}(\eta) = -\mathrm{sgn}(\nu)$, and the second term more specifically for the value $\eta = -\mathrm{sgn}(\nu)$. Upon abbreviating $\Omega(x, \xi, \eta) = \{y \in \Omega | y = C(x, \eta J_s) - J_p m\xi\}$:

$$\begin{aligned}\max_{\xi,\eta\in[-1,1], y\in\Omega(x,\xi,\eta)} L(\xi, \eta, y) &= \max_{\xi,\eta\in[-1,1], y\in\Omega(x,\xi,\eta)} \{\psi(y) + B(y + J_p m\xi, |\eta| J_s) - \nu\eta\} \\ &= \max_{|\xi|,|\eta|\leqslant 1, y\in\Omega(x,\xi,|\eta|)} \{\psi(y) + B(y + J_p m\xi, |\eta| J_s) + |\nu\eta|\},\end{aligned}$$

and

$$\begin{aligned}\max_{\xi,\eta\in[-1,1], y\in\Omega} L(\xi, \eta, y) &= \max_{\xi,\eta\in[-1,1], y\in\Omega} \{\psi(y) + B(y + J_p m\xi, |\eta| J_s) - \nu\eta\} \\ &= \max_{|\xi|\leqslant 1, y\in\Omega} \{\psi(y) + B(y + J_p m\xi, J_s)\} + |\nu|.\end{aligned}$$

We observe the potential consistency of assuming $\psi(x)$ to increase monotonically for $x \geqslant 0$. An increase in x leads via the constraint $y \in \Omega(x, \xi, |\eta|)$ to an increase of y inside the first maximization, so that $\psi(y)$ will increase. The term with $B(.,.)$ will also increase if the sign of ξ is chosen right. So we make the ansatz that $\psi(x)$ is differentiable, and that $\psi'(x) \geqslant 0$ on $x \geqslant 0$. This implies that $\Omega = [-u, u]$, with $\max_{x\in\Omega} \psi(x) = \psi(u) = 0$. The second maximization in (B.30) now reduces to

$$\begin{aligned}\max_{y\in\Omega, |\xi|\leqslant 1} \{\psi(y) + B(y + J_p m\xi, J_s)\} + |\nu| &= \psi(u) + B(u + J_p|m|, J_s) + |\nu| \\ &= B(u + J_p|m|, J_s) + |\nu|.\end{aligned} \tag{B.33}$$

This simplifies our equation (B.30) for $\psi(x)$. For all $x \in [0, u]$ we now have

$$\begin{aligned}\psi(x) &= \max_{|y|\leqslant u, y=C(x,|\eta|J_s)-J_p m\xi, |\xi|,|\eta|\leqslant 1} \{\psi(y) + B(C(x, |\eta|J_s), |\eta|J_s) + |\nu|(|\eta| - 1)\} \\ &\quad - B(u + J_p|m|, J_s), \\ &= \max_{|y|\leqslant u, |y-C(x,|\eta|J_s)|\leqslant J_p|m|, |\eta|\leqslant 1} \{\psi(y) + B(C(x, |\eta|J_s), |\eta|J_s) + |\nu|(|\eta| - 1)\} \\ &\quad - B(u + J_p|m|, J_s), \\ &= \max_{|\eta|\leqslant 1} \max_{y\in[-u,u]\cap[C(x,|\eta|J_s)-J_p|m|, C(x,|\eta|J_s)+J_p|m|]} \\ &\quad \times \{\psi(y) + B(C(x, |\eta|J_s), |\eta|J_s) + |\nu|(|\eta| - 1)\} - B(u + J_p|m|, J_s).\end{aligned} \tag{B.34}$$

Since $\psi(y)$ is monotonic in $|y|$, we need $|y|$ to be as large as possible for any given $|\eta|$. Since the intersection interval (if it exists) is always biased to the right, we must find the largest allowed value y in the intersection interval. The intersection is seen to be empty if $C(x, |\eta|J_s) > u + J_p|m|$, whereas the remaining possible scenarios are

$$\begin{aligned}&u - J_p|m| < C(x, |\eta|J_s) < u + J_p|m| : \quad y_{\max} = u, \\ &C(x, |\eta|J_s) < u - J_p|m| : \quad y_{\max} = C(x, |\eta|J_s) + J_p|m|.\end{aligned}$$

Consistency with the premise $x \in [0, u]$ demands that we must identify the point where x becomes so large that the intersection interval is empty for *any* value of $|\eta|$ should be the boundary $x = u$. This, together with $\min_{|\eta|\leqslant 1} C(x, |\eta|J_s) = C(x, J_s)$, immediately gives us an equation for u: $C(u, J_s) = u + J_p|m|$, or equivalently

$$u = A(u + J_p|m|, J_s). \tag{B.35}$$

Graphical inspection shows that this equation always has one unique non-negative solution u. Within our present construction we can always achieve a non-empty intersection set in (B.34) for suitable (ξ, η), and we may proceed with maximization over $|\eta|$. For each $x \in \Omega$ we now have

$$\begin{aligned}
\psi(x) &= \max_{|\eta|\leqslant 1, C(x,|\eta|J_s)\leqslant u+J_p|m|} \begin{cases} \psi(C(x,|\eta|J_s)+J_p|m|)+B(C(x,|\eta|J_s),|\eta|J_s)+|\nu||\eta| \\ \qquad \text{if} \quad C(x,|\eta|J_s) < u - J_p|m| \\ B(u+J_p|m|,|\eta|J_s)+|\nu||\eta| \\ \qquad \text{if} \quad C(x,|\eta|J_s) > u - J_p|m| \end{cases} \\
&\quad - B(u + J_p|m|, J_s) - |\nu|, \\
&= \max_{z\in[0,J_s], C(x,z)\leqslant u+J_p|m|} \begin{cases} \psi(C(x,z)+J_p|m|)+B(C(x,z),z)+|\nu|z/J_s \\ \qquad \text{if} \quad C(x,z) < u - J_p|m| \\ B(u+J_p|m|,z)+|\nu|z/J_s \\ \qquad \text{if} \quad C(x,z) > u - J_p|m| \end{cases} \\
&\quad - B(u + J_p|m|, J_s) - |\nu|, \\
&= \max_{z\in[C(x,u+J_p|m|),J_s]} \begin{cases} \psi(C(x,z)+J_p|m|)+B(C(x,z),z)+|\nu|z/J_s \\ \qquad \text{if} \quad z > C(x,u - J_p|m|) \\ B(u+J_p|m|,z)+|\nu|z/J_s \\ \qquad \text{if} \quad z < C(x,u - J_p|m|) \end{cases} \\
&\quad - B(u + J_p|m|, J_s) - |\nu|.
\end{aligned} \tag{B.36}$$

Since both $C(x, z)$ and $B(C(x, z), z)$ decrease monotonically with increasing z (see appendix C) we are sure that for sufficiently small values of ν we always find the maximum in (B.36) by substituting the smallest allowed value for z. We now proceed by assuming this property to hold for *any* value of ν. If indeed we always need the smallest z, namely $z = C(x, u + J_p|m|)$, we obtain for all $x \in [0, u]$:

$$\begin{aligned}
\psi(x) &= B(u + J_p|m|, C(x, u + J_p|m|)) - B(u + J_p|m|, J_s) \\
&\quad + \frac{|\nu|C(x, u + J_p|m|)}{J_s} - |\eta|.
\end{aligned} \tag{B.37}$$

This expression meets our requirements: it increases monotonically on $[0, u]$, and (using the general identity $C(x, C(x, y)) = y$ in combination with our previously established relation $C(u, J_s) = u + J_p|m|$) one verifies that $\psi(u) = 0$. We take this as sufficient support for our ansatz; in addition we will find that for the purpose of evaluating the scalar order parameters (m, k) and the phase diagrams we do not need the full shape of $\psi(x)$ but only the property that $\psi(-u) = \psi(u) = \max_{x\in\Omega} \psi(x)$ with $u = A(u + J_p|m|, J_s)$.

What remains in our present analysis is to work out the order parameter equations for m and k, and confirm that these support the premise $\lim_{n\to\infty} k = k_0$. For large but finite n one would expect to have $k = k_0 + k_1/n + \mathcal{O}(n^{-2})$ for $n \to \infty$, which implies that $n\beta\xi(J_p - J_g)(k - k_0) = \beta\xi(J_p - J_g)k_1 + \mathcal{O}(n^{-1})$. Similarly one would expect for large but finite n that $\log \Psi(x) = n\beta\psi(x) + \psi_1(x) + \mathcal{O}(n^{-1})$, with $\psi(x)$ as given by (B.37). Insertion of

these forms into (81) and (82) gives integrals over (x, y) that can be evaluated by the steepest descent, with the relevant saddle point obtained for $x = y = u\,\mathrm{sgn}(m\xi)$:

$$\begin{aligned} m &= \lim_{n\to\infty} \\ &\quad \times \frac{\int \mathrm{d}\xi\; w(\xi)\xi \int_{-u}^{u} \mathrm{d}x\,\mathrm{d}y\,\tanh[\beta(J_p m\xi + x + y)]\,\mathrm{e}^{n[\log\cosh[\beta(J_p m\xi+x+y)]+\beta\psi(x)+\beta\psi(y)]+\psi_1(x)+\psi_1(y)+\beta\xi(J_p-J_g)k_1}}{\int \mathrm{d}\xi\; w(\xi) \int_{-u}^{u} \mathrm{d}x\,\mathrm{d}y\,\mathrm{e}^{n[\log\cosh[\beta(J_p m\xi+x+y)]+\beta\psi(x)+\beta\psi(y)]+\psi_1(x)+\psi_1(y)+\beta\xi(J_p-J_g)k_1}}, \\ &= \lim_{n\to\infty} \frac{\int \mathrm{d}\xi\; w(\xi)|\xi|\,\mathrm{sgn}(m)\,\tanh[\beta(J_p|m\xi|+2u)]\,\mathrm{e}^{n\log\cosh[\beta(J_p|m\xi|+2u)]+2\psi_1(u\,\mathrm{sgn}(m\xi))+\beta\xi(J_p-J_g)k_1}}{\int \mathrm{d}\xi\; w(\xi)\,\mathrm{e}^{n\log\cosh[\beta(J_p|m\xi|+2u)]+2\psi_1(u\,\mathrm{sgn}(m\xi))+\beta\xi(J_p-J_g)k_1}}, \end{aligned} \tag{B.38}$$

so

$$|m| = \tanh[\beta(J_p|m| + 2u)]. \tag{B.39}$$

Similarly we must solve

$$\begin{aligned} k_0 &= \lim_{n\to\infty} \frac{\int \mathrm{d}\xi\; w(\xi)\xi \int_{-u}^{u} \mathrm{d}x\,\mathrm{d}y\,\mathrm{e}^{n[\log\cosh[\beta(J_p m\xi+x+y)]+\beta\psi(x)+\beta\psi(y)]+\psi_1(x)+\psi_1(y)+\beta\xi(J_p-J_g)k_1}}{\int \mathrm{d}\xi\; w(\xi) \int_{-u}^{u} \mathrm{d}x\,\mathrm{d}y\,\mathrm{e}^{n[\log\cosh[\beta(J_p m\xi+x+y)]+\beta\psi(x)+\beta\psi(y)]+\psi_1(x)+\psi_1(y)+\beta\xi(J_p-J_g)k_1}}, \\ &= \lim_{n\to\infty} \frac{\int \mathrm{d}\xi\; w(\xi)\xi\,\mathrm{e}^{n\log\cosh[\beta(J_p|m\xi|+2u)]+2\psi_1(u\,\mathrm{sgn}(m\xi))+\beta\xi(J_p-J_g)k_1}}{\int \mathrm{d}\xi\; w(\xi)\,\mathrm{e}^{n\log\cosh[\beta(J_p|m\xi|+2u)]+2\psi_1(u\,\mathrm{sgn}(m\xi))+\beta\xi(J_p-J_g)k_1}}, \\ &= \frac{\mathrm{e}^{2\psi_1(u\,\mathrm{sgn}(m))+\beta(J_p-J_g)k_1} - \mathrm{e}^{2\psi_1(-u\,\mathrm{sgn}(m))-\beta(J_p-J_g)k_1}}{\mathrm{e}^{2\psi_1(u\,\mathrm{sgn}(m))+\beta(J_p-J_g)k_1} + \mathrm{e}^{2\psi_1(-u\,\mathrm{sgn}(m))-\beta(J_p-J_g)k_1}}. \end{aligned} \tag{B.40}$$

As soon as a solution for the non-leading order $\psi_1(x)$ exists, there will be a value of k_1 that gives the desired value $k = k_0$. However, careful inspection of the sub-leading orders in the functional saddle-point equation for $\Psi(x)$ reveals that the above construction works for $\nu < 0$, but no finite solution $\psi_1(x)$ exists when $\nu > 0$. In the latter case it turns out that the solution of the problem scales with n as $\log\Psi(x) = \beta n\psi(x) + \psi_1(x)\sqrt{n} + \mathcal{O}(n^0)$ and $k = k_0 + k_1/\sqrt{n} + \cdots$. For a detailed analysis of the different sub-leading orders see appendix D.

The final result is that $k = k_0$ solutions always exist (although they will be locally stable only for $J_g > J_p$), and that the associate value of the order parameter m is to be solved from the two coupled equations

$$|m| = \tanh[\beta(2u + J_p|m|)], \tag{B.41}$$

$$\tanh(\beta u) = \tanh[\beta(u + J_p|m|)\tanh(\beta J_s). \tag{B.42}$$

The sign of m is arbitrary, both solutions $m = \pm|m|$ are allowed and equally likely. We solve the first equation for u, giving $u = \frac{1}{2}\beta^{-1}\tanh^{-1}(|m|) - \frac{1}{2}J_p|m|$, and obtain an equation involving $|m|$ only:

$$\frac{\tanh\left[\frac{1}{2}\tanh^{-1}(|m|) - \frac{1}{2}\beta J_p|m|\right]}{\tanh\left[\frac{1}{2}\tanh^{-1}(|m|) + \frac{1}{2}\beta J_p|m|\right]} = \tanh(\beta J_s). \tag{B.43}$$

The joint distribution $W(h, \xi)$ of effective fields and polarities for the present solution, where $J_p m \neq 0$, is found to be

$$\begin{aligned} W(h,\xi) &= \tfrac{1}{2}(1 + k_0)\delta(\xi - 1)\delta[h - J_p m - 2u\,\mathrm{sgn}(m)] \\ &\quad + \tfrac{1}{2}(1 - k_0)\delta(\xi + 1)\delta[h + J_p m + 2u\,\mathrm{sgn}(m)] \end{aligned} \tag{B.44}$$

The free energy per monomer (B.5) for the present type of solution is found to reduce to

$$\lim_{n\to\infty}\varphi = \tfrac{1}{2}J_p\left(m^2 + k_0^2\right) - \tfrac{1}{2}J_g\left(k_0^2 - k^{\star 2}\right) - |\nu| - B\left(\tfrac{1}{2}\beta^{-1}\tanh^{-1}(|m|) + \tfrac{1}{2}J_p|m|,\, J_s\right). \tag{B.45}$$

Appendix C. Properties of the functions $C(x,y)$ and $B(C(x,y),y)$

The functions $B(x, y)$ and $C(x, y)$ are defined as

$$B(x, y) = \frac{1}{2\beta} \log[4 \cosh[\beta(x + y)] \cosh[\beta(x - y)]], \tag{C.1}$$

$$C(x, y) = \beta^{-1} \tanh^{-1}[\tanh(\beta x)/ \tanh(\beta y)]. \tag{C.2}$$

We are only interested in the regime where $y \geqslant 0$ and $|x| < y$. The function $C(x, y)$ is monotonic and anti-symmetric in x, and obeys $\mathrm{sgn}[C(x, y)] = \mathrm{sgn}(xy)$ and $|C(x, y)| \geqslant |x|$. It is the x inverse of $A(x, y)$, since

$$\begin{aligned} A(C(x, y), y) &= \beta^{-1} \tanh^{-1}[\tanh(\beta C(x, y)) \tanh(\beta y)] \\ &= \beta^{-1} \tanh^{-1}[\tanh(\beta x)] = x, \\ C(A(x, y), y) &= \beta^{-1} \tanh^{-1}[\tanh(\beta A(x, y))/ \tanh(\beta y)] \\ &= \beta^{-1} \tanh^{-1}[\tanh(\beta x)] = x. \end{aligned}$$

Furthermore $C(x, y)$ obeys the general identity

$$\begin{aligned} C(x, C(x, y)) &= \beta^{-1} \tanh^{-1} \left[\frac{\tanh(\beta x)}{\tanh(\beta x)/ \tanh(\beta y)} \right] \\ &= \beta^{-1} \tanh^{-1}[\tanh(\beta y)] = y. \end{aligned} \tag{C.3}$$

The function $B(x, y)$ is symmetric in x; thus also the function $B(C(x, y), y)$ is symmetric in x. The partial derivatives of $C(x, y)$ are

$$\frac{\partial}{\partial x} C(x, y) = \frac{\tanh(\beta y)[1 - \tanh^2(\beta x)]}{\tanh^2(\beta y) - \tanh^2(\beta x)}, \tag{C.4}$$

$$\frac{\partial}{\partial y} C(x, y) = -\frac{\tanh(\beta x)[1 - \tanh^2(\beta y)]}{\tanh^2(\beta y) - \tanh^2(\beta x)}. \tag{C.5}$$

Next we work out and simplify the quantity $B(C(x, y), y)$ with the help of identities such as

$$2 \cosh[\tanh^{-1}(m) + \beta y] = \mathrm{e}^{\beta y} \left(\frac{1 + m}{1 - m} \right)^{\frac{1}{2}} + \mathrm{e}^{-\beta y} \left(\frac{1 + m}{1 - m} \right)^{-\frac{1}{2}}$$

$$\begin{aligned} &2 \cosh \left[\tanh^{-1} \left(\frac{\tanh(\beta x)}{\tanh(\beta y)} \right) + \beta y \right] \cosh \left[\tanh^{-1} \left(\frac{\tanh(\beta x)}{\tanh(\beta y)} \right) - \beta y \right] \\ &\qquad = \frac{\tanh^2(\beta y) + \tanh^2(\beta x)}{\tanh^2(\beta y) - \tanh^2(\beta x)} + \cosh(2\beta y). \end{aligned}$$

This results in

$$\begin{aligned} B(C(x, y), y) &= \frac{1}{2\beta} \log \left\{ 4 \cosh \left(\tanh^{-1} \left[\frac{\tanh(\beta x)}{\tanh(\beta y)} \right] + \beta y \right) \right. \\ &\qquad \left. \times \cosh \left(\tanh^{-1} \left[\frac{\tanh(\beta x)}{\tanh(\beta y)} \right] - \beta y \right) \right\} \\ &= \frac{1}{\beta} \log[2 \cosh(\beta y)] - \frac{1}{\beta} \log \cosh(\beta x) - \frac{1}{2\beta} \log \left[1 - \frac{\tanh^2(\beta x)}{\tanh^2(\beta y)} \right]. \end{aligned} \tag{C.6}$$

Hence we have

$$\frac{\partial}{\partial x}B(C(x,y),y)=\frac{\tanh(\beta x)[1-\tanh^2(\beta y)]}{\tanh^2(\beta y)-\tanh^2(\beta x)}. \tag{C.7}$$

Thus, in the region $|x|<|y|$ we have $\frac{\partial}{\partial x}B(C(x,y),y)<0$ for $x<0$ and $\frac{\partial}{\partial x}B(C(x,y),y)>0$ for $x>0$. The function $B(C(x,y),y)$ is symmetric in x, diverges at $x=\pm y$, and has a unique minimum $B(C(0,y),y)=\beta^{-1}\log[2\cosh(\beta y)]$ at $x=0$.

Appendix D. Analysis of sub-leading orders for the state $k=k_0$ as $n\to\infty$

Here we analyze in more detail the sub-leading terms in n of the non-trivial solution of our equations (80)–(82) for the case where $J_g>J_p$, i.e. where $m\neq 0$ and $k=k_0$, as $n\to\infty$. Given the exponential scaling with n of the kernel in (80), we may without loss of generality for $n\to\infty$ always write $\Psi(x)$ in one of the following two forms:

$$\text{either}: \quad \Psi(x)=\mathrm{e}^{n\psi(x)+\psi_1(x)+\mathcal{O}(n^{-1})}, \tag{D.1}$$

$$\text{or}: \quad \Psi(x)=\mathrm{e}^{n\psi(x)+\sqrt{n}\psi_1(x)+\mathcal{O}(n^0)}. \tag{D.2}$$

Since $\psi(x)$ was found to be maximal at $x=\pm u$ (where $u>0$), we find in both cases

$$\lim_{n\to\infty}\Psi(x)=\alpha\delta(x-u)+(1-\alpha)\delta(x+u), \tag{D.3}$$

where

$$\text{scaling (D.1)}: \qquad \alpha=\frac{\mathrm{e}^{\psi_1(u)}}{\mathrm{e}^{\psi_1(u)}+\mathrm{e}^{\psi_1(-u)}}, \tag{D.4}$$

$$\text{scaling (D.2)}: \qquad \alpha=\theta[\psi_1(u)-\psi_1(-u)]. \tag{D.5}$$

We will show below that for $\nu<0$ the solution is of the form (D.1), with $k=k_0+k_1/n+\cdots$,

$$k_1=0, \qquad \alpha=\frac{\sqrt{1+\mathrm{sgn}(m)k_0}(\sqrt{1+|k_0|}-\sqrt{1-|k_0|})}{2|k_0|}, \tag{D.6}$$

and with $\lim_{n\to\infty}p(\xi)=w(\xi)$, whereas for $\nu>0$ the solution is of the form (D.2), with $k=k_0+k_1/\sqrt{n}+\cdots$,

$$k_1=\frac{\psi_1(-u)-\psi_1(u)}{\beta\,\mathrm{sgn}(m)(J_p-J_g)}, \qquad \alpha=\theta[\psi_1(u)-\psi_1(-u)], \tag{D.7}$$

and with $\lim_{n\to\infty}p(\xi)=\delta[\xi+\mathrm{sgn}(k_1)]$.

D.1. First scaling ansatz: $\mathcal{O}(n^0)$ sub-leading terms

If we simply substitute (D.3) and $k=k_0+k_1/n+\cdots$ into equation (80), we find

$$\lim_{n\to\infty}p(\xi)=\frac{w(\xi)\,\mathrm{e}^{\beta\xi(J_p-J_g)k_1}}{\int\mathrm{d}\xi' w(\xi')\,\mathrm{e}^{\beta\xi'(J_p-J_g)k_1}}, \tag{D.8}$$

and

$$\begin{aligned}
&\alpha\delta(x-u)+(1-\alpha)\delta(x+u)\\
&\quad=\lim_{n\to\infty}\frac{\alpha\int\mathrm{d}\xi\,\mathrm{d}\eta\,p(\xi)w(\eta)\delta[x-A(J_pm\xi+u,\eta J_s)]\,\mathrm{e}^{n\beta[B(J_pm\xi+u,\eta J_s)-\nu\eta]}}{\int\mathrm{d}\xi\,\mathrm{d}\eta\,p(\xi)w(\eta)\{\alpha\,\mathrm{e}^{n\beta[B(J_pm\xi+u,\eta J_s)-\nu\eta]}+(1-\alpha)\,\mathrm{e}^{n\beta[B(J_pm\xi-u,\eta J_s)-\nu\eta]}\}}\\
&\qquad+\lim_{n\to\infty}\frac{(1-\alpha)\int\mathrm{d}\xi\,\mathrm{d}\eta\,p(\xi)w(\eta)\delta[x-A(J_pm\xi-u,\eta J_s)]\,\mathrm{e}^{n\beta[B(J_pm\xi-u,\eta J_s)-\nu\eta]}}{\int\mathrm{d}\xi\,\mathrm{d}\eta\,p(\xi)w(\eta)\{\alpha\,\mathrm{e}^{n\beta[B(J_pm\xi+u,\eta J_s)-\nu\eta]}+(1-\alpha)\,\mathrm{e}^{n\beta[B(J_pm\xi-u,\eta J_s)-\nu\eta]}\}}.
\end{aligned} \tag{D.9}$$

Since $\eta \in [-1, 1]$ and $B(.,.)$ is symmetric and monotonically increasing in both arguments, the leading exponentials are maximal for $\eta = -\mathrm{sgn}(\nu)$ and $\xi = \pm\mathrm{sgn}(m)$, so

$$\alpha\delta(x-u)+(1-\alpha)\delta(x+u) = \frac{\alpha p(\mathrm{sgn}(m))\delta[x+\mathrm{sgn}(\nu)A(J_p|m|+u,J_s)]+(1-\alpha)p(-\mathrm{sgn}(m))\delta[x-\mathrm{sgn}(\nu)A(J_p|m|+u,J_s)]}{\alpha p(\mathrm{sgn}(m))+(1-\alpha)p(-\mathrm{sgn}(m))}. \tag{D.10}$$

There are two possibilities for solution, dependent on how we match the two δ-peaks on either side of this equation. One always ends up with u to be solved from

$$u = A(J_p|m|+u, J_s), \tag{D.11}$$

but, since $u > 0$, the specific matching depends on ν. For $\nu < 0$ one is forced to choose

$$\alpha = \frac{\alpha\, \mathrm{e}^{\beta\mathrm{sgn}(m)(J_p-J_g)k_1}}{\alpha\, \mathrm{e}^{\beta\mathrm{sgn}(m)(J_p-J_g)k_1}+(1-\alpha)\,\mathrm{e}^{-\beta\mathrm{sgn}(m)(J_p-J_g)k_1}}, \tag{D.12}$$

whereas for $\nu > 0$ the only option is

$$\alpha = \frac{(1-\alpha)\mathrm{e}^{-\beta\mathrm{sgn}(m)(J_p-J_g)k_1}}{\alpha\, \mathrm{e}^{\beta\mathrm{sgn}(m)(J_p-J_g)k_1}+(1-\alpha)\,\mathrm{e}^{-\beta\mathrm{sgn}(m)(J_p-J_g)k_1}}. \tag{D.13}$$

To proceed with equations (81) and (82) for m and k we first calculate

$$\begin{aligned}\int \mathrm{d}\xi\, W(h,\xi)\xi f(h) &= \lim_{n\to\infty}\frac{\int \mathrm{d}\xi\,\mathrm{d}x\,\mathrm{d}y\, w(\xi)\,\mathrm{e}^{\beta\xi(J_p-J_g)k_1}\Psi(x)\Psi(y)f(x+y+J_pm\xi)\xi\,\mathrm{e}^{n\log\cosh[\beta(x+y+J_pm\xi)]}}{\int \mathrm{d}\xi\,\mathrm{d}x\,\mathrm{d}y\, w(\xi)\,\mathrm{e}^{\beta\xi(J_p-J_g)k_1}\Psi(x)\Psi(y)\,\mathrm{e}^{n\log\cosh[\beta(x+y+J_pm\xi)]}},\\ &= \mathrm{sgn}(m)\frac{\alpha^2 f(2u+J_p|m|)\,\mathrm{e}^{\beta\mathrm{sgn}(m)(J_p-J_g)k_1}-(1-\alpha)^2 f(-2u-J_p|m|)\,\mathrm{e}^{-\beta\mathrm{sgn}(m)(J_p-J_g)k_1}}{\alpha^2\,\mathrm{e}^{\beta\mathrm{sgn}(m)(J_p-J_g)k_1}+(1-\alpha)^2\,\mathrm{e}^{-\beta\mathrm{sgn}(m)(J_p-J_g)k_1}}.\end{aligned} \tag{D.14}$$

Application of this formula to $f(h) = \tanh(\beta h)$ and $f(h) = 1$ gives

$$|m| = \tanh[\beta(2u+J_p|m|)], \tag{D.15}$$

$$k_0 = \mathrm{sgn}(m)\frac{\alpha^2\,\mathrm{e}^{\beta\mathrm{sgn}(m)(J_p-J_g)k_1}-(1-\alpha)^2\,\mathrm{e}^{-\beta\mathrm{sgn}(m)(J_p-J_g)k_1}}{\alpha^2\,\mathrm{e}^{\beta\mathrm{sgn}(m)(J_p-J_g)k_1}+(1-\alpha)^2\,\mathrm{e}^{-\beta\mathrm{sgn}(m)(J_p-J_g)k_1}} \tag{D.16}$$

So far we have successfully recovered the equations for u and m are as derived earlier; the next question is whether we can find a corresponding solution for k_1 and α.

Both (D.12) and (D.13) are quadratic equations for α, so we expect at most two solutions. In fact for $\nu > 0$ only one of these is in the interval $[0, 1]$:

$$\nu < 0: \qquad \alpha \in \{0, 1\}, \tag{D.17}$$

$$\nu > 0: \qquad \alpha = \frac{1}{1+\mathrm{e}^{\beta\mathrm{sgn}(m)(J_p-J_g)k_1}}, \tag{D.18}$$

For $\nu < 0$, combination with (D.4) and (D.16) subsequently gives

$$k_1 = 0, \qquad \alpha = \frac{\sqrt{1+\mathrm{sgn}(m)k_0}\left(\sqrt{1+|k_0|}-\sqrt{1-|k_0|}\right)}{2|k_0|}. \tag{D.19}$$

For $\nu > 0$, on the other hand, the solution breaks down. Upon writing k_1 in terms of α and substituting the result into (D.16), we find the trivial $k_0 = 0$. Thus, only for the degenerate special case $k_0 = 0$ is the solution of our equations for $\nu > 0$ of the form (D.1). We conclude that the generic solution for $\nu > 0$ scales differently with n.

D.2. Second scaling ansatz: $\mathcal{O}(\sqrt{n})$ sub-leading terms

If we substitute (D.3) and $k = k_0 + k_1/\sqrt{n} + \cdots$ into equation (80) (where $k_1 \neq 0$, since otherwise we return to the previous scaling case) we get

$$\lim_{n\to\infty} p(\xi) = \delta[\xi + \mathrm{sgn}(k_1)] \tag{D.20}$$

and

$$\begin{aligned}
&\alpha\delta(x-u) + (1-\alpha)\delta(x+u) \\
&= \lim_{n\to\infty} \frac{\alpha \int \mathrm{d}\xi\, \mathrm{d}\eta\, p(\xi)w(\eta)\delta[x - A(J_p m\xi + u, \eta J_s)]\, \mathrm{e}^{n\beta[B(J_p m\xi+u,\eta J_s)-\nu\eta]}}{\int \mathrm{d}\xi\, \mathrm{d}\eta\, p(\xi)w(\eta)\left\{\alpha\, \mathrm{e}^{n\beta[B(J_p m\xi+u,\eta J_s)-\nu\eta]} + (1-\alpha)\, \mathrm{e}^{n\beta[B(J_p m\xi-u,\eta J_s)-\nu\eta]}\right\}} \\
&\quad + \lim_{n\to\infty} \frac{(1-\alpha) \int \mathrm{d}\xi\, \mathrm{d}\eta\, p(\xi)w(\eta)\delta[x - A(J_p m\xi - u, \eta J_s)]\, \mathrm{e}^{n\beta[B(J_p m\xi-u,\eta J_s)-\nu\eta]}}{\int \mathrm{d}\xi\, \mathrm{d}\eta\, p(\xi)w(\eta)\left\{\alpha\, \mathrm{e}^{n\beta[B(J_p m\xi+u,\eta J_s)-\nu\eta]} + (1-\alpha)\, \mathrm{e}^{n\beta[B(J_p m\xi-u,\eta J_s)-\nu\eta]}\right\}}.
\end{aligned} \tag{D.21}$$

Once more the dominant exponent is maximal when $\eta = -\mathrm{sgn}(\nu)$ and $\xi = \pm\mathrm{sgn}(m)$, so

$$\begin{aligned}
&\alpha\delta(x-u) + (1-\alpha)\delta(x+u) \\
&\quad = \lim_{n\to\infty} \frac{\mathrm{e}^{\sqrt{n}[\psi_1(u)+\beta\mathrm{sgn}(m)(J_p-J_g)k_1]}\delta[x + \mathrm{sgn}(\nu)A(J_p|m|+u, J_s)]}{\mathrm{e}^{\sqrt{n}[\psi_1(u)+\beta\mathrm{sgn}(m)(J_p-J_g)k_1]} + \mathrm{e}^{\sqrt{n}[\psi_1(-u)-\beta\mathrm{sgn}(m)(J_p-J_g)k_1]}} \\
&\qquad + \lim_{n\to\infty} \frac{\mathrm{e}^{\sqrt{n}[\psi_1(-u)-\beta\mathrm{sgn}(m)(J_p-J_g)k_1]}\delta[x - \mathrm{sgn}(\nu)A(J_p|m|+u, J_s)]}{\mathrm{e}^{\sqrt{n}[\psi_1(u)+\beta\mathrm{sgn}(m)(J_p-J_g)k_1]} + \mathrm{e}^{\sqrt{n}[\psi_1(-u)-\beta\mathrm{sgn}(m)(J_p-J_g)k_1]}}.
\end{aligned} \tag{D.22}$$

Again we have to match the two δ-peaks on both sides. Since we know that the equation $u = -A(J_p|m|+u, J_s)$ has no non-negative solutions u (for $J_p m \neq 0$), we are forced to match $\delta(x \pm u)$ to $\delta[x \pm A(J_p|m|+u, J_s)]$. From this we recover equation (D.11), as required, but now with

$$\nu < 0: \quad \alpha = \lim_{n\to\infty} \frac{\mathrm{e}^{\sqrt{n}[\psi_1(u)+\beta\mathrm{sgn}(m)(J_p-J_g)k_1]}}{\mathrm{e}^{\sqrt{n}[\psi_1(u)+\beta\mathrm{sgn}(m)(J_p-J_g)k_1]} + \mathrm{e}^{\sqrt{n}[\psi_1(-u)-\beta\mathrm{sgn}(m)(J_p-J_g)k_1]}}, \tag{D.23}$$

$$\nu > 0: \quad \alpha = \lim_{n\to\infty} \frac{\mathrm{e}^{\sqrt{n}[\psi_1(-u)-\beta\mathrm{sgn}(m)(J_p-J_g)k_1]}}{\mathrm{e}^{\sqrt{n}[\psi_1(u)+\beta\mathrm{sgn}(m)(J_p-J_g)k_1]} + \mathrm{e}^{\sqrt{n}[\psi_1(-u)-\beta\mathrm{sgn}(m)(J_p-J_g)k_1]}}. \tag{D.24}$$

Our present equations can be obtained from those of the previous scaling regime upon substituting $k_1 \to \sqrt{n}k_1$ and $\psi_1(x) \to \sqrt{n}\psi_1(x)$. This allows us to take over the previous evaluation of the order parameter equations for m and k, provided we make the appropriate substitutions. For m we then recover equation (D.11) (as required), whereas the equation for k gives

$$k_0\,\mathrm{sgn}(m) = \lim_{n\to\infty} \frac{\mathrm{e}^{\sqrt{n}[2\psi_1(u)+\beta\mathrm{sgn}(m)(J_p-J_g)k_1]} - \mathrm{e}^{\sqrt{n}[2\psi_1(-u)-\beta\mathrm{sgn}(m)(J_p-J_g)k_1]}}{\mathrm{e}^{\sqrt{n}[2\psi_1(u)+\beta\mathrm{sgn}(m)(J_p-J_g)k_1]} + \mathrm{e}^{\sqrt{n}[2\psi_1(-u)-\beta\mathrm{sgn}(m)(J_p-J_g)k_1]}}. \tag{D.25}$$

We have now successfully recovered the expressions for u and m derived earlier; the remaining question is whether we can find a corresponding solution for k_1 and α from the coupled equations (D.5), (D.23)–(D.25). Since $|k_0| < 1$ we conclude from (D.25) that the following must be true, so that the $\mathcal{O}(\sqrt{n})$ terms cancel and the $\mathcal{O}(n^0)$ terms can indeed give us $|k_0| < 1$:

$$k_1 = \frac{\psi_1(-u) - \psi_1(u)}{\beta\mathrm{sgn}(m)(J_p - J_g)}. \tag{D.26}$$

This solution for k_1 we can insert into our previous equations for α, which gives

$$\nu < 0: \quad \alpha = \lim_{n\to\infty} \frac{\mathrm{e}^{\sqrt{n}\psi_1(-u)}}{\mathrm{e}^{\sqrt{n}\psi_1(-u)} + \mathrm{e}^{\sqrt{n}\psi_1(u)}} = 1 - \alpha, \tag{D.27}$$

$$\nu > 0: \quad \alpha = \lim_{n\to\infty} \frac{e^{\sqrt{n}\psi_1(u)}}{e^{\sqrt{n}\psi_1(-u)} + e^{\sqrt{n}\psi_1(u)}} = \alpha. \tag{D.28}$$

Apparently, for $\nu > 0$ the present scaling ansatz gives self-consistent solutions. For $\nu < 0$ we find $\alpha = \frac{1}{2}$, and hence $k_1 = 0$ which is forbidden since it effectively brings us back to the previous scaling regime. We conclude that, apart from degenerate limits, the two scaling ansatz (D.1) and (D.2) are complementary: for $\nu < 0$ the system is in a state of the type (D.1), whereas for $\nu > 0$ it is in a state of the type (D.2).

References

[1] Guo W, Shea J E and Berry R S 2005 *Ann. New York Acad. Sci.* **1066** 34–53
[2] Echenique P 2007 *Contemp. Phys.* **48** 81–108
[3] Anfinsen C B 1973 *Science* **181** 223–30
[4] Daggett V 2006 *Chem. Rev.* **106** 1898–916
[5] Bryngelson J D and Wolynes P G 1987 *Proc. Natl Acad. Sci. USA* **84** 7524–8
[6] Derrida B 1981 *Phys. Rev.* B **24** 2613–26
[7] Lau K F and Dill K A 1989 *Macromolecules* **22** 3986–97
[8] Prentiss M C, Hardin C, Eastwood M P, Zong C and Wolynes P G 2006 *J. Chem. Theory Comput.* **2** 705–16
[9] Chen N-Y, Su Z-Y and Mou C-Y 2006 *Phys. Rev. Lett.* **96** 078103
[10] Yang J S, Chen W W, Skolnick J and Shakhnovich E I 2007 *Structure* **15** 53–63
[11] Aktürk E, Arkin H and Celik T 2007 *Preprint* cond.mat/0703606
[12] Onuchic J N, Luthey-Schulten Z and Wolynes P G 1997 *Annu. Rev. Chem.* **48** 545–600
[13] Pande V S and Rokhsar D S 1999 *Proc. Natl Acad. Sci USA* **96** 9062–7
[14] Faccioli P, Sega M, Pederiva F and Orland H 2006 *Phys. Rev. Lett.* **97** 108101
[15] Dill K A 1984 *Biochemistry* **24** 1501–9
[16] Dill K A 1990 *Biochemistry* **29** 7133–55
[17] Kauzmann W 1959 *Adv. Protein Chem.* **14** 1–63
[18] Rose G D and Wolfenden R 1993 *Annu. Rev. Biophys. Struct.* **22** 381–415
[19] Sippi M J 1996 *J. Mol. Biol.* **260** 644–8
[20] Rose G, Flemming P, Banavar J R and Maritan A 2006 *Proc. Natl Acad. Sci. USA* **103** 16623–33
[21] Collet O 2005 *Europhys. Lett.* **72** 301–7
[22] Oberdorf R, Ferguson A, Jacobsen J L and Kondev A 2006 *Phys. Rev.* E **74** 051801
[23] Abkevich V I, Gutin A M and Shakhnovich E I 1994 *J. Chem. Phys.* **101** 6052–62
[24] Kenzaki H and Kikuchi M 2006 *Chem. Phys. Lett.* **427** 414–7
[25] Das P, Moll M, Stamati H, Kavraki L E and Clementi C 2006 *Proc. Natl Acad. Sci. USA* **103** 9885–90
[26] Skantzos N S, van Mourik J and Coolen A C C 2001 *J. Phys. A: Math. Gen.* **34** 4437–57
[27] Bascle J, Garel T and Orland H 1993 *J. Phys. I France* **3** 259–75
[28] Konkoli Z, Hertz J and Franz S 2001 *Phys. Rev.* E **64** 051910
[29] Wilder J and Shakhnovitch E I 2000 *Phys. Rev.* E **62** 7100–10
[30] Müller M, Mézard M and Montanari A 2004 *J. Chem. Phys.* **120** 11233–55
[31] Coolen A C C, Penney R W and Sherrington D 1993 *Phys. Rev.* B **48** 16116–8
[32] Penney R W, Coolen A C C and Sherrington D 1993 *J. Phys. A: Math. Gen.* **26** 3681–95
[33] Jongen G, Bollé D and Coolen A C C 1998 *J. Phys. A: Math. Gen.* **31** L737–42
[34] Jongen G, Anemüller J, Bollé D, Coolen A C C and Pérez-Vicente C J 2001 *J. Phys. A: Math. Gen.* **34** 3957–84
[35] Chakravorty H, Coolen A C C and Sherrington D 2002 *J. Phys. A: Math. Gen.* **35** 8647–71
[36] Nikoletopoulos T and Coolen A C C 2004 *J. Phys. A: Math. Gen.* **37** 8433–56
[37] Nikoletopoulos T, Coolen A C C, Pérez-Castillo I, Skantzos N S, Hatchett J P L and Wemmenhove B 2004 *J. Phys. A: Math. Gen.* **37** 6455–75
[38] Hatchett J P L, Skantzos N S and Nikoletopoulos T 2005 *Phys. Rev.* E **72** 066105
[39] Heylen R, Skantzos N S, Busquets Blanco J and Bollé D 2006 *Phys. Rev.* E **73** 016138
[40] Parker J M R 1999 *J. Comput. Chem.* **20** 947–55
[41] Murzin A G, Brenner S E, Hubbard T and Chothia C 1995 *J. Mol. Biol.* **247** 536–40
[42] Andreeva A, Howorth D, Chandonia J M, Brenner S E, Hubbard T J, Chothia C and Murzin A G 2008 *Nucl. Acids. Res.* **36** D419–25
[43] Eisenberg D, Schwarz E, Komarony M and Wall R 1984 *J. Mol. Biol.* **179** 125–42
[44] Moelbert S, Emberly E and Tang C 2004 *Protein Sci.* **13** 752–62