Scuola Internazionale Superiore di Studi Avanzati - Trieste

SISSA ISAS

Thesis submitted for the degree of
Doctor Philosophiæ

Statistical Physics and Message Passing Algorithms. Two Case Studies: MAX-K-SAT Problem and Protein Flexibility

Candidate: Michal Kolář
Supervisors: Prof. Riccardo Zecchina
Prof. Cristian Micheletti

Trieste, December 2005

SISSA - Via Beirut 2-4 - 34014 TRIESTE - ITALY

Preface

In the last decades the theory of spin glasses has been developed within the framework of statistical physics. The obtained results showed to be novel not only from the physical point of view, but they have brought also new mathematical techniques and algorithmic approaches. Indeed, the problem of finding ground state of a spin glass is (in general) NP-complete.

The methods that were found brought new ideas to the field of Combinatorial Optimization, and on the other side, the similar methods of Combinatorial Optimization, were applied in physical systems. As it happened with the Monte Carlo sampling and the Simulated Annealing, also the novel Cavity Method lead to algorithms that are open to wide use in various fields of research. The Cavity Method shows to be equivalent to Bethe Approximation in its most symmetric version, and the derived algorithm is equivalent to the Belief Propagation, an inference method used widely for example in the field of Pattern Recognition.

The Cavity Method in a less symmetric situation, when one has to consider correctly the clustering of the configuration space, lead to a novel message–passing algorithm—the Survey Propagation.

The class of Message–Passing algorithms, among which both the Belief Propagation and the Survey Propagation belong, has found its application as Inference Algorithms in many engineering fields. Among others let us mention the Low–Density Parity–Check Codes, that are widely used as Error–Correcting Codes for communication over noisy channels.

In the first part of this work we have compared efficiency of the Survey Propagation Algorithm and of standard heuristic algorithms in the case of the random-MAX-K-SAT problem. The results showed that the algorithms perform similarly in the regions where the clustering of configuration space does not appear, but that the Survey Propagation finds much better solutions to the optimization problem in the critical region where one has to consider existence of many ergodic components explicitly.

The second part of the thesis targets the problem of protein structure and flexibility. In many proteins the mobility of certain regions and rigidity of other regions of their structure is crucial for their function or interaction with other cellular elements. Our simple model tries to point out the flexible regions from the knowledge of native 3D–structure of the protein. The problem is mapped to a spin glass model which is successfully solved by the Believe Propagation algorithm.

Contents

Chapter 1

Introduction

We would like to show algorithmic approaches inspired by methods of statistical physics and to show their efficiency when compared with the standard methods of combinatorial optimization.

In the first part of the thesis we will study satisfiability problem, the first–to–know NP-complete problem. The latter part will consider the structure and flexibility of proteins.

Although the two problems, one from the field of combinatorial optimization and the other from the field of theoretical biophysics, do not seem to be related straightforwardly, they share the property of being defined on a dilute graph and hence being in general solvable by methods of modern statistical physics in conjuncture with algorithms borrowed from the field of combinatorial optimization. Furthermore, both of them can be reformulated as constrained satisfaction problems.

In our work we apply general algorithms to solve the two problems and we derive new algorithms specifically adapted for them. We also compare the acquired results with outcome of well established methods.

In computer science the field of combinatorial optimization [15] deals with the general issue of classifying the computational complexity ("hardness") of decision or optimization problems and of designing search algorithms. Similarly to statistical physics models, a generic combinatorial optimization problem is composed of many discrete variables—*e.g.*, Boolean variables, finite sets of colours, or Ising spins and p-state spins—which interact through constraints typically involving a small number of variables, that in turn sum up to give the global cost-energy function.

1.1 The Complexity Theory

Before we start with devising the methods to solve the problems, it will be useful to summarize the main results of the complexity theory and to state the difference between the worst–case and the average–case complexity.

To start we will define what we mean by the *problem*, its *instance*, *solution* and *algorithmic complexity*.

By a problem we mean any general question to be answered. In the case of our work the questions will be: *Is there any satisfying assignment for the given logical formula?*, *What is the smallest number of violated clauses of that formula?*, and in the part where we will speak about proteins it will be the question *How flexible is a given three-dimensional structure?*.

Already when asking these questions, we have separated implicitly the problem and its instance; Instance is the specific realization of the problem which is defined by a concrete specification of all free parameters of the problem. Hence we ask: *Is there a satisfying assignment for the formula A?*, *What is the smallest number of violated clauses in the formula B?*, and finally *What is the number of floppy modes in the structure C?*. Here A, B, and C define instances of three different problems.

It is our nature to try to answer any question posed. The answer to a question we will call a solution. Yet there are various solutions we can give. To the so called *decision problems* we answer only *yes* or *no*. An example would be the question *Is there a satisfiable assignment?*. To other questions we have to give a more detailed answer. Thus the solution of *What is the smallest number of violated clauses?* will be a number. The same is true for the question that asks for the number of floppy modes of a mechanical structure. The problems that demand also for some solution are called *search problem*. The answer to an *optimization problem* is then the best answer that one is able to give together with the appropriate assignment of variables, that is the free parameters of the model.

For finding the solution of a problem we may use our intuition, but sometimes we prefer to have a stepwise procedure that would give at least some solution for any instance of the problem—an *algorithm* solving the problem.

If we know the algorithm solving a problem we are sure that we can, in some time, find the solution of any instance of the given problem. Still we do not know how much time we will spend searching for the solution. There may be faster and more efficient ways and, on the other side, there may be algorithms that are just too inefficient to be used in practice.

The time demand of the algorithm will certainly depend on a given instance of the problem, so we have to find a measure of the instance difficulty, its size. We will call *instance size* the amount of data that defines the instance uniquely but economically. This we do with a coding of the instance (there are, certainly good and bad codings, see [34] for a detailed description of the coding). The size of the instance is then given by the number of letters of a good coding of the instance.

Once knowing what is the size of the instance, we can measure how scales the time needed by the algorithm to provide an answer with the instance size. Still there are many instances of a given size, so we have to specify which instance we will consider as the example of all instances of that size

N. The two most wide–spread approaches are to choose either the worst case instances or the average case instances. The *worst–case complexity* then gives the time needed by the algorithm in the worst combination of input parameters for the given size N of the instance and hence it bounds the time complexity from above. There are many problems in which the worst case instances turn out to be atypical and one does not meet them in common praxis. Then the worst–case complexity looses its meaning of a time estimator and it is much more convenient to work with the *average–case complexity* that states how long it will *usually* take to obtain the result.

As was conjectured during many years of research, there are generally two classes of problems. Those problems for which there exists an algorithm which worst–case complexity is bounded polynomially in N, and the algorithms for which the time–complexity grows faster than any polynomial. Such heavy–to–deal–with problems are called *intractable*. Let us remark that even if there is only a single instance of the problem that cannot be solved in polynomial time, the problem is still considered as intractable in the means of worst–case complexity.

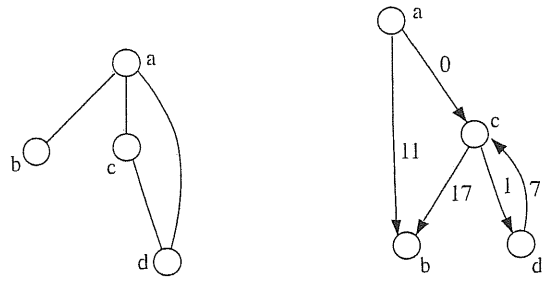
The first class of the problems is called P (P for polynomial). The other class, *intractable problems*, can be still split into two parts. The problems whose solution can be guessed and then verified in polynomial time, and the rest. The first part of the problems actually forms a class called *NP* (non-deterministic polynomial). A class of equivalence is defined as a set of all problems that can be mapped to each other through an algorithmic transformation which complexity is at most polynomial.

1.1.1 The NP-Completeness

Are all the *NP*-problems equally difficult or is there any subset of that class that may be considered as the hardest core? Are there any problems that, once we know algorithms that solve them, can be mapped to other problems in a fast way in means of some (polynomial) reduction? The answer is yes, there are *NP* problems that can be polynomially transformed to each other and any other *NP* problem can be reduced to some of them. They are called *NP-complete*.

Already from their definition we see that the NP-complete problems form an equivalency class with respect to a polynomial transformation. The first problem that was shown to belong to this class was the satisfiability problem [2].

Before we start with detailed description of the two problems that we want to examine, let us summarize in brief some basics of the graph theory. It will show extremely useful when finding reasonable codings for our problems.



a) b)

Figure 1.1: **a)** A simple graph with four nodes and four links. bf b) A simple weighted directed graph.

1.2 Basic Definitions of the Graph Theory

In all systems that we are going to talk about, the easiest way how to represent a given instance of the problem is to map the instance to some graph. The structure of these graphs depends on the problem we speak about, but they certainly share some general properties. Let us summarize them in the following paragraphs.

We start with the definition of a *graph*. A graph $g(V, E)$ is composed of a set of nodes (vertices) V and a set of edges E. Each edge connects two vertices, on the other side, there may be many edges sharing a single vertex.

The edges (links) may or may not be oriented and the graph is then called *directed* or not-oriented, respectively. One can also place on each edge a number, a weight, the graph is then called weighted.

An example of a graph with weighted links would be the graph representing the road map of Trieste where many roads are one-way. The crossings would be vertices of the graph and the streets its links. The weights would correspond to distances between crossroads.

We will call *degree* of the vertex the number of links connected to the vertex. So for example node a in the Figure 1.1a has degree equal to 3 while the degree of the node b is 1. It is also reasonable to distinguish the *in-degree* of a node, that is the number of oriented links entering it and the *out-degree* of a node, that is the number of outgoing links. As an example take node c of the directed graph in the Figure 1.1b, it has the in-degree equal to 2. Similarly the out-degree of the node a is 2.

We will call node b to be *neighbour* of a node a if there is an edge $(a, b) \in E$. We will use notation $b \in g(a)$ for the node b neighbouring the node a in the graph $g(V, E)$. Any given graph is uniquely defined by specifying the set of neighbours for each node.

There are several properties of graphs that we can measure and use for characterization of the graphs. For overall description of the graph we use the connectivity distribution, and also the clustering coefficient which provides information about the coarse grained structure of the graph, it may point out existence of hub nodes (nodes with large connectivity and a 'central' role in the graph). The presence of loops and their length may be crucial for convergence of our methods. We will see that existence of short loops may cause impossibility to find solution. Fortunately enough, the graphs that we will meet will have either small number of short loops, or we will be able to use a method that is not critically sensitive to their existence.

In the presented work we will be interested in two kinds of graphs. First, in the case of random K-SAT model, we will have random graphs with randomly distributed connectivity of the nodes. In a sense structure of this graph is very similar to the structure of randomly generated graphs of the Erdős–Renyi type.

In the part of the thesis where we will talk about the protein structure and flexibility, we will use graphs that directly reflect the real three dimensional structure of the protein.

1.2.1 Erdős-Renyi Random Graphs

As there are many graphs that come up by coding of various instances of a combinatorial optimization problem, we may be interested to know typical properties of the solution of the problem on some typical instance. Many times we may such a typical case represent by a random graph that maintains only the characteristic properties of the problem. Sometimes one may continue in the abstraction even further and study the problem properties on a completely random structure. This is very close to a study of typical complexity of the problem, rather than worst-case complexity.

The most straightforward way how to generate a random graph is to take some number of vertices and to place an edge with some probability p between any pair of vertices a, b. We may create the complete ensemble of graphs of that kind and to study its properties. This ensemble takes names of Erdős and Renyi [3] and although the graphs of the ensemble have been created in a random way, we may still find some general characteristics of this ensemble.

As was already told, for a real world problem, we can find the connectivity distribution of the graph that codes an instance of the problem. We may, hence, be interested in the connectivity distribution of the Erdős-Renyi graphs. If the distribution of the real–world–derived graph is similar to the one of random graphs, we can use random graphs as a good trial ensemble for algorithmic studies.

For the Erdős–Renyi random graphs the connectivity distribution follows the binomial distribution and in the limit of large yet dilute graphs Poisson

distribution. Indeed the probability to find a node with k neighbours is:

$$\text{Prob}(k) = \binom{N-1}{k} p^k (1-p)^{N-1-k}, \tag{1.1}$$

which in the limit of large N and fixed pN goes to:

$$\text{Prob}(k) = \frac{(pN)^k}{k!} e^{-pN}. \tag{1.2}$$

1.2.2 Factor Graphs

Special case of graphs that will be used in the following work are *factor graphs*. In these graphs the vertices are divided into two different sets, and the edges are present only between nodes of different kind. A simple factor graph is shown in the Figure 1.6. This kind of graphs can be conveniently used in constraint satisfaction problems: one kind of nodes may represent the constraints applied to the variables, which are represented as the other type of nodes.

1.3 Spin Models on Diluted Structures

Many problems of combinatorial optimization can be straightforwardly mapped to a physical problem defined on a factor graph. It comes from the nature of combinatorial problems, that each constraint of the problem contains only relatively small number of variables, but the complete cost function, the energy in the physical point of view, is a sum of many such constraints.

This immediately implies that the graph that represents the interactions of the variables is rather dilute, although the total number of vertices may be huge. The degree of any node will be relatively small and the number of edges will scale approximately linearly with the number of vertices. Physical system corresponding to that kind of cost function is the *spin glass*.

Physical motivation to study similar problems come from studies of magnetic alloys. Imagine that one has a sample of dia– or paramagnetic metal and in this material there is a very small addition of ferromagnetic material. The atoms of ferromagnetic metal are randomly distributed in the alloy. The magnetic coupling between the atoms strongly oscillates with the distance so we can expect the coupling values to be distributed almost randomly in some range.

For the energy of such a spin system we can write a toy–model Hamiltonian:

$$H = \sum_{contacts} J_{ij} s_i s_j, \tag{1.3}$$

where the actual realization of the couplings $\{J_{ij}\}$ between atoms i and j depends on the sample that we have prepared. Mean–field theories for the

glassy Hamiltonian were studied. In the Sherington–Kirkpatrick model, the couplings are statistically uncorrelated and their values are distributed according to some distribution with zero mean. The actual form of the distribution is believed not to be crucially important, as a reflection of the universality phenomenon. Two simplest and most studied versions are Gaussian distributed couplings and couplings distributed to be 1 or -1 uniformly. Mean–field theories for spin glasses defined on a dilute graph are also available.

Since we are interested in finding any physical properties that will be true for any sample we make, we have to average over the couplings somehow. It is a useful abstraction to consider the couplings as another type of variables and to express the Hamiltonian as a function of the original spins $\{s_i\}$ and the couplings $\{J_{ij}\}$, $H(\{s_i\}, \{J_{ij}\})$.

However, there is a fundamental difference between the two sets of variables. While for any sample the values of $\{J_{ij}\}$ are fixed, the spin variables are free to flip and relax according to the current values $\{J_{ij}\}$. The dynamics of the couplings have to be much slower and we will treat them as some kind of *quenched variables*.

In order to calculate average properties of a spin glass we have to calculate given properties for a single sample and then average this quantities over ensemble of samples, all possible realizations of the quenched variables.

$$\ll \mathcal{O} \gg = \sum_{\{J_{ij}\}} \text{Prob}(\{J_{ij}\}) \langle \mathcal{O} \rangle_{\{J_{ij}\}}, \tag{1.4}$$

where $\langle \mathcal{O} \rangle_{\{J_{ij}\}}$ is the expectation value of the observable $\mathcal{O}$ given the configuration of couplings $\{J_{ij}\}$.

The expectation value $\langle \mathcal{O} \rangle_{\{J_{ij}\}}$ is the customary statistical average over canonical ensemble:

$$\langle \mathcal{O} \rangle_{\{J_{ij}\}} = \frac{1}{Z} \sum_{\{s_i\}} e^{-\beta H(\{J_{ij}\},\{s_i\})} \mathcal{O}. \tag{1.5}$$

Where Z is the partition function of the glassy Hamiltonian:

$$Z = \sum_{\{s_i\}} e^{-\beta H(\{J_{ij}\},\{s_i\})}. \tag{1.6}$$

1.3.1 The Replica Method

Certainly we cannot calculate the spin configurations and hence the value of the observable $\langle \mathcal{O} \rangle$ for each sample. We have to calculate the mean value $\ll \mathcal{O} \gg$ directly using some trick. The replica trick.

Having the partition function of the system (1.6), we can introduce a system that contains many samples (replicas) of the original alloy, say n of them.

The complete partition function will be

$$Z^n = \prod_{a=1}^{n} \sum_{\{s_i^a\}} e^{-\beta H(\{J_{ij}\},\{s_i^a\})}. \tag{1.7}$$

The samples a differ only in the realization of spin orientations. If we consider the limit $n \to \infty$ we expect that the observable $\mathcal{O}$ averaged using this partition function will directly give $\ll \mathcal{O} \gg$. Indeed, we do a kind of averaging over all possible configurations of the spins.

More precisely, for the replicated system we can define averaged partition function in the form

$$Z^{(n)} = \langle Z^n \rangle_{\{J_{ij}\}}, \tag{1.8}$$

and using the following limit equality

$$\begin{aligned} x - 1 &= \ln x, \quad x \to 1, \\ Z^n - 1 &= n \ln Z, \quad n \to 0, \end{aligned} \tag{1.9}$$

we may write for the free energy averaged over realizations of the couplings:

$$\langle f \rangle_{\{J_{ij}\}} = \langle \ln Z \rangle_{\{J_{ij}\}} = \lim_{n \to 0} \left\langle \frac{Z^n - 1}{n} \right\rangle_{\{J_{ij}\}} = \lim_{n \to 0} \frac{Z^{(n)} - 1}{n}. \tag{1.10}$$

The free energy obtained within this scheme gives exact solution if we are able to prove existence of the analytical extension of the function $Z^{(n)}$ in $n \in \mathbf{R}^+$. That is something we can do simply for high temperatures by expanding $Z^{(n)}$ in powers of β. It will be on the other side nontrivial in the low temperature regime of the spin glass.

Indeed, in the high temperature regime we expect that the particular assignment of the values J_{ij} does not matter that much, thermal energy of spins s_i will be very high.

When decreasing the temperature, the actual configuration of quenched variables will matter more and more, we will see details of the model. So we have to expect separation of the configuration space in several (or many) distinct ergodic components. This is caused by very rough energy landscape that is observed in the spin glasses as a result of frustration and disorder.

The analogy with a lake where the water level corresponds to energy is quite illustrative and will help us to understand what is happening at small temperatures, see Figure 1.2.

The scenario of the level drop depends on the roughness of the energy landscape. While in the *RS* (replica symmetric, we will see later what we mean by that) case the configuration space shrinks, in the *RSB* (replica symmetry broken) systems the configuration space clusters at some energy. This is illustrated in the Figure 1.2-1RSB. With decreasing temperature, the energy in our system decreases and so the water level in our illustration goes down.

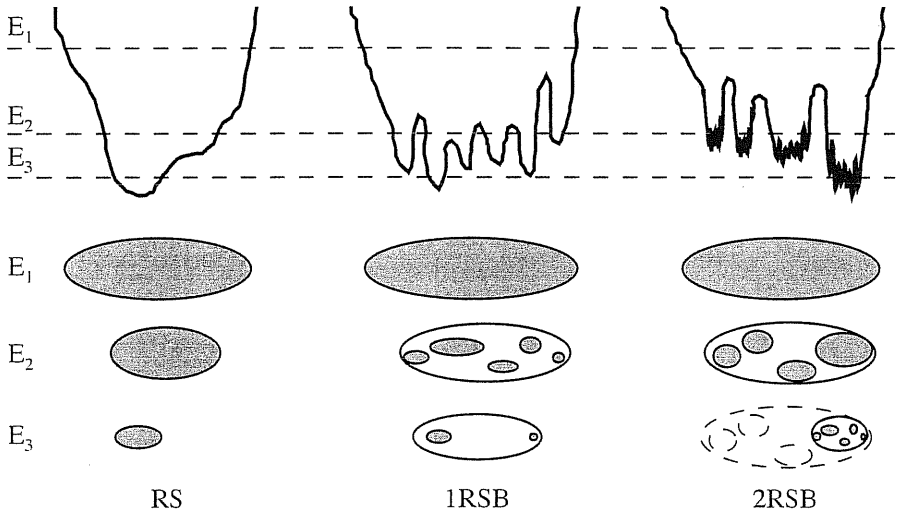


Figure 1.2: The lake illustration of energy landscapes. **RS)** Replica symmetric landscape: no clustering occurs at any energetic level. **1RSB)** First step of replica symmetry breaking. The configurations already form a simple hierarchy with states that can be observed at the energy E_2. **2RSB)** The energy landscape is more rough and the hierarchy has two levels. The states emerge at the energy E_2 and then they are again clustered in new states at energy E_3.

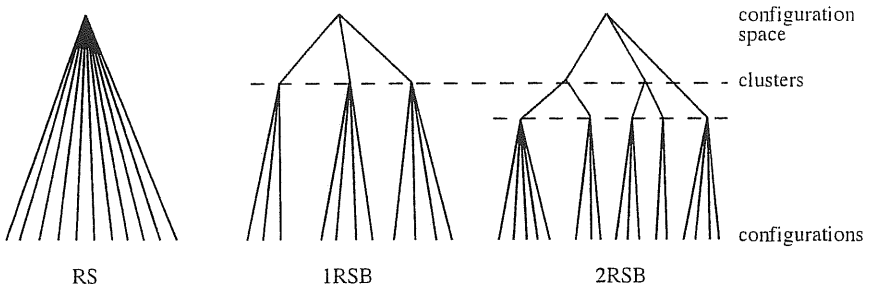


Figure 1.3: The trees of similarities of replicas in replica symmetric (RS), one step replica symmetry broken (1RSB) and two steps replica symmetry broken system. Compare with the Figure 1.2.

As for the spin glasses the lake bed resemble in its roughness the landscape of Carso, at some temperature (corresponding to E_2 in our figure) the configuration space (the water–level) separates in several well distinguished clusters, states. With decreasing energy the area of the allowed regions shrinks and at very low energy, eventually, there is no configuration with the given energy.

The replicas of the system are not equivalent, if we sketch a hierarchy of their similarities—overlaps, it will have a branching. With increasing replica asymmetry the number of branching levels will increase, see Figure 1.3. Eventually, we reach the full–replica–symmetry–broken system in which there are infinite levels of branching.

There are natural problems leading to hierarchical trees that have two and more branching levels. The overlaps of configurations hence have three or more well separated values.

The two case problems that we are going to study are either RS (protein structure) and $RS/1RSB/fullRSB$ in the case of satisfiability problem. What are the reasons of the configuration space clustering?

1.3.2 Disorder & Frustration

In order to understand better the emergence of complex organization of the configuration space in the spin glasses, we will discuss two phenomena—disorder at the level of interactions and at the level of topology, and consecutive frustration.

Let us start with the first spin model that one encounters which is usually the Ising model for the ferromagnet or model of anti-ferromagnet on a regular lattice. The spins tend to align in the direction parallel (anti-parallel) to the neighbouring spins. Ferromagnetic system at low temperature shows two ergodic components—*states*—that differ by a number of spin flips that is of order of the size of the system. Indeed at low enough temperature (below

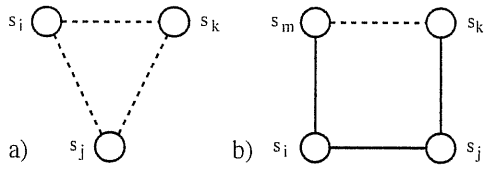


Figure 1.4: Emergence of frustration in system **a)** with even couplings but topological disorder and **b)** with even structure but couplings distributed unevenly. The ferromagnetic coupling are shown as full lines, the dashed lines represent anti-ferromagnetic couplings.

Curie temperature) almost all spins point in the same direction, what in the Ising representation means that the spins point almost all down or up.

However the existence of multiple states is not the only feature that we can observe in spin systems, although it has important practical consequences as we will see in the case of satisfiability problem. The other feature of great importance, that makes solution of some problems quite a difficult task is *frustration*.

Imagine an anti-ferromagnetic system where the spins are placed on some structure in which loops of an odd length may occur. In Figure 1.4a we show a triangular loop with three anti-ferromagnetic spins. When trying to assign values to spins s_i, s_j, and s_k trying to minimize the energy function $\sum_{bonds} \delta_{s_m,s_n}$, we can start with the spin s_i and make it $s_i = \uparrow$. Then we can assign $s_j = \downarrow$ according to the anti-ferromagnetic coupling of the link (ij). The problem arises when we want to assign the value of the spin s_k. Both $s_k = \downarrow$ and $s_k = \uparrow$ will increase the energy because of the existence of one neighbour that is oriented in parallel. Hence the decision of orientation of s_k cannot be based on local information only, we have to use also possible external influences, *i.e.*, other spins of the system in contact with i, j, k, and hence the non-local information. We have to consider the system as a whole which definitely makes computations more complex.

As we have seen, the problem of frustration occurs when there are two concurrent forcings of the spin that have similar strength. In the example of anti-ferromagnetic spins the cause of the frustration was topological. We would not see the situation on a regular rectangular lattice, there must be some disorder that causes irregularity of the lattice: a vacancy, a doping.

On the other side, we can expect the frustration to occur also on a regular lattice where disorder is introduced by uneven assignment of couplings, see Figure 1.4.

Among the tools for numerical investigations of such complex systems at low temperatures the simulated annealing (SA) algorithm [14] and its variants have played a major role. Stochastic processes of this kind satisfy de-

tailed balance and their behaviour can be compared with static and dynamical mean-field calculations. However, in problems in which the interest is focused on zero temperature ground states and where the proliferation of metastable states causes an exponential slowdown in the equilibration rate, the applicability of SA-like algorithms is limited to relatively small system sizes.

In the last few years there has been a great progress in the study of spin glasses over random graphs which has shed new light on mean-field theory and has produced new algorithmic tools for the study of low energy states in large single problem instances. Quite surprisingly, problems which were considered to be algorithmically hard for local search algorithms, like for instance random K-SAT close to the phase boundary, turned out to be efficiently solved by the Survey Propagation algorithm arising from the replica symmetry broken approach to diluted spin glasses.

1.3.3 The Cavity Method

The replica technique is a great theoretical device and its introduction has lead to conceptual findings of the configuration space clustering. On the other side, its application to single instances of problems of statistical physics is rather tricky and, furthermore, there is lack of mathematical theory behind of it.

On the other side, the current results show, that there is another method of statistical physics that is equivalent to the replica method and in which we may state all the assumptions explicitly. The *cavity method* [26, 24] proceeds in the following way.

As the topological structure we work on is a diluted graph, we can virtually remove one spin from the graph together with all the couplings in which it takes part. Furthermore, because the graph is diluted and its structure is at least in some sense random, we expect that it contains only very small number of short loops and hence the spins that are well apart will have uncorrelated values of spin s_i, at least in the replica symmetric case when all the configurations belong to the same pure state.

If we remove a site from such a graph, we know that only its neighbours will be affected directly, and all the other spins only indirectly, the introduced change will cause only small changes in the overall energy. It will be illustrative to show what we mean by this on the computation of the energy of the system. For simplicity, we will work at zero temperature, and only after we will devise formulae also for finite temperature [26].

We can imagine the graph as a growing structure. Starting from a single spin, we can add all the other spins one–by–one together with the interaction links. If we know, how to calculate the energy of $n + 1$ spins from the energy of n spins, we may recursively calculate complete energy of the system. Still, the graph structure changes dramatically when we start from $n = 1$ and grow

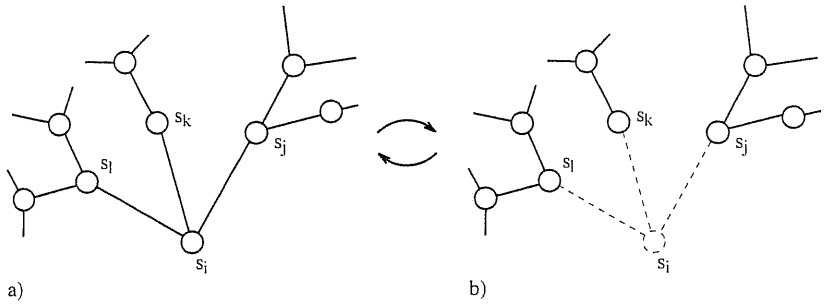


Figure 1.5: Removal of the site s_i creates cavity in the graph. The rest of the tree–like graph is separated into three parts that we can deal 'separately'. Sites s_j, s_k, and s_l may be considered uncorrelated in the graph with cavity **b)**, since we suppose that there are no short loops in the full graph **a)**.

the graph to its actual size. Hence we can imagine another operation, which will lead to coupled consistency equations for the energy. The operation will be removal of a single site, together with its links, and after relaxation, its reintroduction into the graph. During this operation we create a cavity in the graph and it is this cavity that has given the name to the method.

Hence our procedure will be the following: We remove a randomly chosen site s_i from the graph, we let the system relax, what at the zero temperature means that we will minimize the energy, and then we return the spin back to the system. After relaxation we will be able to express the energy E^N of the complete graph as some function of energy of the graph with the cavity E_i^{N-1}. See the Figure 1.5 for illustration.

The necessary condition which allows us to do so, is the separation of the Hamiltonian in a sum over contacts of the graph. That is possible for our spin glass. The topological consequence of this property is that the graph may be redrawn in a form of a bipartite graph, which is called a factor graph, see Figure 1.6. In our case, the spin Hamiltonian obeys the condition, it is a sum over link contributions $J_{ij}s_is_j$. The Hamiltonian of the system without the site i may be written in the form

$$H^{N-1} = \sum_{m<n;\ m,n\neq i} J_{mn}s_ms_n, \tag{1.11}$$

After thermal relaxation (at zero temperature) the energy reaches its minimum

$$E^{N-1} = \min_{\{s_m\}} \sum_{m<n;\ m,n\neq i} J_{mn}s_ms_n. \tag{1.12}$$

When putting back the site i, we have to include in the Hamiltonian also all

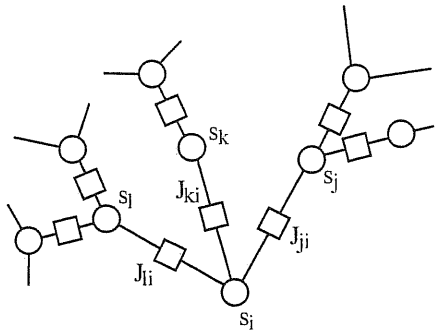


Figure 1.6: The factor graph representation of the spin glass Hamiltonian (1.3). The graph contains two sets of nodes. The circles represent the spin variables s_i, while the squares represent the couplings of the Hamiltonian J_{ij}.

its interactions

$$H^N = H^{N-1} + \sum_{m \in i} J_{mi} s_m s_i, \tag{1.13}$$

where by the notation $m \in i$ we mean all neighbours of the site i in the interaction graph. Letting the system relax we get the energy of the N spin system

$$E^N = \min_{\{s_m\}} H^N. \tag{1.14}$$

Under the assumption of the tree–like structure of the graph, we expect that only a small number of short loops is present in the graph and hence that the variables s_j, s_k, and s_l in the graph with cavity are uncorrelated. The energy can be rewritten in the form of some constant part that depends on the configuration of all the other spins and the three spins in some local fields that are uncorrelated (the abstraction of the fields can be done always, since the spins are Boolean variables, but the fields are uncorrelated only in the structures without the loops)

$$E^{N-1} = E_0^i + \sum_{m \in i} h_m s_m. \tag{1.15}$$

After addition of the spin s_i, all the couplings are turned on, and they add to the energy of the system. After thermal relaxation

$$E^N = E_0^i + \sum_{m \in i} \min_{s_m = \pm 1} \{h_m s_m + J_{mi} s_m s_i\}. \tag{1.16}$$

As the spin s_i has only two possible values, we can write the minimum in the form of some linear function

$$\min_{s_m = \pm 1} \{h_m s_m + J_{mi} s_m s_i\} = -w_{mi} - u_{mi} s_i. \tag{1.17}$$

The term w_{mi} gives some constant energetic contribution and the factor u_{mi} can be considered as a kind of a field that we will call *cavity bias*. Note that the cavity bias has a flavour of direction, it is a field imposed in the cavity of the variable s_i by the graph segment connected to it through the variable s_m. And it is actually this property that gives idea of using the cavity biases as messages that we will use in the algorithmic way later on.

Finally we can express the energy of the system with the help of cavity biases in the form

$$E^N = E_0^i + \sum_{m \in i} \left[-w_{mi} - u_{mi} s_i \right], \tag{1.18}$$

where the cavity bias u and the constant factor w can be expressed as

$$\begin{aligned} u_{mi} &= \operatorname{sgn}(h_m J_{mi}) \min \{|J_{mi}|, |h_m|\}, \\ w_{mi} &= \operatorname{sgn}(h_m J_{mi}) \max \{|J_{mi}|, |h_m|\}, \end{aligned} \tag{1.19}$$

as can be shown by simple enumeration of all possible cases.

This formula may be applied to any site i of the graph, hence we obtain a set of coupled equations that can be used for evaluation of the energy in the self–consistent way.

At a finite temperature we cannot evaluate the energy directly, we have to start rather with the probability distribution of a configuration $\{s_i\}$. This probabilistic interpretation will bring us closer to the field of combinatorial optimization and to the message passing algorithms.

We start again with the cavity on the site of the spin s_i, as shown in the Figure 1.5. Let, for example, the spins s_j, s_k, and s_l be the only neighbours of the spin s_i in the graph. As we suppose that the graph representing the Hamiltonian is diluted and it has a tree–like structure, so the eventual loops are very long[1], we can consider that the spins neighbouring to s_i are uncorrelated in the graph with cavity. Indeed, either they lie in different subtrees of the graph, and they are completely uncorrelated, or they have very large distance so we can consider them uncorrelated. The graph is virtually separated into three subtrees.

The probability of finding a configuration of the three spins $\{s_j, s_k, s_l\}$ is hence factorized

$$P^{N-1}(s_j, s_k, s_l) \simeq P_j^{N-1}(s_j) P_k^{N-1}(s_k) P_l^{N-1}(s_l). \tag{1.20}$$

If we return the spin s_i to the graph together with all its couplings, we can calculate probability of any configuration of the quadruplet of the spins.

$$P^N(s_i, s_j, s_k, s_l) = \prod_{m \in i} \left[P_m^{N-1}(s_m) e^{-\beta J_{im} s_i s_m} \right], \tag{1.21}$$

[1] As it is true for random graphs, for which it was shown [3] that the shortest loops that are macroscopically present in the graph have length of order of logarithm of the number of vertices.

where we have not forgotten to place the Boltzmann factor corresponding to the couplings of s_i to its neighbours. Once knowing the probability distributions of spin values in the cavity graph we can calculate the complete probability distribution in means of some self–consistent procedure.

To make the similarity with the energy calculation at zero energy temperature straightforward, we may write the probability $P_m^{N-1}(s_m)$ in the form of Boltzmann factor of the spin s_m in some effective field

$$P_m^{N-1}(s_m) = \frac{e^{-\beta h_m s_m}}{2\cosh(\beta h_m)}. \tag{1.22}$$

We will call the field h_m the *cavity field*. Similarly for the variable s_i we may write

$$P_i^N(s_i) = \frac{e^{-\beta h_i s_i}}{2\cosh(\beta h_i)}, \tag{1.23}$$

what we can certainly do, as there are only two states of the spin s_i.

Moreover, we can find the probability $P_i^N(s_i)$ also by marginalizing the probability (1.21) over the neighbours of s_i,

$$P^N(s_i) = \prod_{m\in i} \sum_{s_m=\pm 1} \left[P_m^{N-1}(s_m) e^{-\beta J_{im} s_i s_m} \right]. \tag{1.24}$$

The inner sums may be expressed in the way similar to (1.23)

$$\sum_{s_m=\pm 1} \left[P_m^{N-1}(s_m) e^{-\beta J_{im} s_i s_m} \right] = C_{mi} e^{\beta u_{mi} s_i}, \tag{1.25}$$

where the normalization factor C_{mi} and the *cavity bias* u_{mi} can be expressed in a closed form similarly to the equation (1.19). The probability of finding the spin i in the state s_i is then

$$P^N(s_i) = \prod_{m\in i} C_{mi} \quad e^{\beta \sum_{m\in i} u_{mi} s_i}. \tag{1.26}$$

Let us imagine that the structure we work with is random, and so the spin Hamiltonian. That means generally that the couplings J_{ij} are randomly drawn from some distribution. In this case we may compare the equations (1.23) and (1.26).

We expect that the cavity fields h_m, that we have defined in (1.22), are also random variables chosen from some distribution $\mathcal{P}_m(h_m)$. In the random graph and in the thermodynamic limit we expect that all the spins will be equivalent, and hence that there is some probability distribution $\mathcal{P}(h)$ such that $\mathcal{P}_m(h_m) = \mathcal{P}(h_m)$ for any m. Moreover, since the spin s_i is in nothing special to other spins, we expect $\mathcal{P}_i(h_i) = \mathcal{P}(h_i)$ either. Similarly to $\mathcal{P}(h)$ we may define the distribution $\mathcal{Q}(u)$ of the cavity biases. Comparing (1.23) and (1.26) we have

$$h_i = \sum_{m\in i} u_{mi}, \tag{1.27}$$

and hence for the probability distributions

$$P(h_i) = \sum_k \mathrm{Prob}(k) \int \prod_{m\in i} [Q(u_{mi} \mathrm{d}u_{mi})] \delta(h_i - \sum_{m\in i} u_{mi}), \qquad (1.28)$$
$$Q(u_{ij}) = \int \mathcal{P}(h_i) \delta(u_{ij} - \mathrm{sgn}(h_i + J_{ij}) \min\{|h_m|, |J_{ij}|\}).$$

We have introduced the connectivity distribution of the random graph $\mathrm{Prob}(k)$, because although each coupling has only two interacting spins, the degrees of various spins may differ. For the random graph we know exact form of the distribution (1.2).

This set of coupled equations will be central for our calculations, it actually gives replica–symmetric solution to the spin glass problem in the mean–field approximation. The symmetry is introduced through the assumption that the probability distributions are equal for all the spins and that the spins of vertices next to cavity are uncorrelated. This is certainly true only in the case when the configuration space is not clustered, only if there is a single *pure state*. In the problems where the clustering occurs we have to find some more appropriate way how to describe the probability distributions, [1].

It is worth mentioning that these equations are equivalent to the Belief Propagation, a message passing algorithm that was devised in the field of the artificial intelligence—pattern recognition.

Although these equations are valid only if we consider a single cluster we can device a method that will go beyond this restriction. To deal with a number of pure states, we will do a 'survey' of probability distribution within them. The result of a 'survey' will be a distribution functional of probability distributions for h and u. If we expect to find many states we move from a 'histogram' to a functional measure on the space of possible probability distributions $P(h)$ and $Q(u)$. Following this pathway will lead us to the novel Survey Propagation algorithm which solves the problem on the level of the one–step replica breaking.

General Belief Propagation

The Belief Propagation (BP) is an algorithm for inference that was repeatedly invented in the fields of artificial intelligence, pattern recognition, communication technology, and other fields of computer science. We have seen now, that also in statistical physics we may use it for solving problems. Actually, the transfer matrix method is a Belief–Propagation–like algorithm.

One of the prominent applications of the BP is the pattern recognition in artificial vision. The goal of the recognition is to infer the scene s that we are observing and we have its image r. Having the image means that we know some properties r_i of the image (that may be properties of single pixels or of some patches of the pixels). We would like to know if the image properties

correspond to some scene properties s_i. We assume that there is a statistical dependence of properties s_i and r_i (some expert advise) that we denote $\phi_i(s_i, r_i)$. Similarly, we suppose that there is some statistical correlation of scene properties s_i and s_j, a compatibility function $\psi_{ij}(s_i, s_j)$.

The joint probability $P(\{s_i\}, \{r_i\})$ of observing some scene s and having its image r is

$$P(\{s_i\}, \{r_i\}) = \frac{1}{Z} \prod_{(ij)} \psi_{ij}(s_i, s_j) \prod_k \phi_k(s_k, r_k), \tag{1.29}$$

where the first product is done over neighbouring pairs of properties. The normalization constant Z is actually the partition function.

We see immediately the similarity with the spin–glass equations; taking s_i to be spin value, $\psi_{ij}(s_i, s_j) = \exp\{-\beta J_{ij} s_i s_j\}$ and $\phi_k(s_k, r_k) = 1$ or $\phi_k(s_k, r_k) = \exp -\beta h s_k$ in the case of the spin glass in the external field. But let us continue in the derivation of the message passing algorithm in its most general way.

We will define the message $m_{ji}(s_i)$ that will have the same dimensionality as the property s_i and it will express how likely the node j thinks that the node i will be in the state s_i. The *belief* at a node i then will be the product of all incoming messages m_{ji} and of the local 'advice' $\phi_i(s_i, r_i)$:

$$b_i(s_i) = C_i \prod_{j \in i} m_{ji}(s_i) \phi_i(s_i, r_i), \tag{1.30}$$

where C_i is a normalization constant that ascertains that the believes sum up to 1. For simplification, we will suppose that the scene and the image does not change in time and we will write $\phi_i(s_i)$ instead of $\phi_i(s_i, r_i)$.

The messages are to be calculated self–consistently through the update formula

$$m_{ij}(s_j) = \sum_{s_i} \prod_{k \in i \setminus j} m_{ki}(s_i) \psi_{ij}(s_i, s_j) \phi_i(s_i), \tag{1.31}$$

in which we immediately recognize equivalent of the formula (3.19).

If there are no loops in the graph of interactions of properties, the believes calculated from the messages after reaching self–consistence are actually the marginal probabilities for the property i to have value s_i:

$$P_i(s_i) \sim b_i(s_i). \tag{1.32}$$

To demonstrate this, we will show only simple example of Yedidia [57]. First we express the belief at node i in the graph of Figure 1.7 using the definition of belief (1.30)

$$b_i(s_i) = C_i m_{ji}(s_i) \phi_i(s_i). \tag{1.33}$$

Now we apply the message update rule (1.31) and we replace m_{ji} by the messages entering to the node j,

$$b_i(s_i) = C_i \phi_i(s_i) \sum_{s_j} m_{kj}(s_k) m_{lj}(s_l) \psi_{ij} i(s_i, s_j) \phi_j(s_j), \tag{1.34}$$

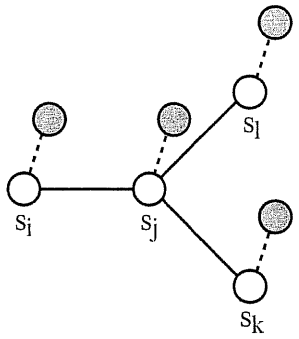


Figure 1.7: A small graph that will help us to understand relation between believes and marginal probabilities. The dark nodes correspond to variables r_i of the image, we are trying to guess the properties of the scene $\{s_i\}$.

and writing explicitly the messages m_{kj} and m_{lj} we have

$$b_i(s_i) = C_i\phi_i(s_i)\sum_{s_j}\sum_{s_k}\psi_{kj}(s_k,s_j)\phi_k(s_k)\sum_{lj}\psi_{lj}(s_l,s_j)\phi_k(s_l)\psi_{ij}i(s_i,s_j)\phi_j(s_j). \tag{1.35}$$

After reorganizing the sums we have

$$b_i(s_i) = C_i\sum_{\{s_j,s_k,s_l\}}\prod_m \phi_m(s_m)\prod_{(mn)}\psi_{mn}(s_m,s_n), \tag{1.36}$$

in what we immediately recognize the joint probability (1.29) marginalized over all properties but i. Hence we have shown that the formula (1.32) is correct in our example.

For our analysis it will be helpful to define also two–point believes $b_{ij}(s_i,s_j)$ that after the convergence of messages to the fixed point, will correspond to the probability marginalized over all properties but i and j.

$$b_{ij}(s_i,s_j) = C_{ij}\psi_{ij}(s_i,s_j)\phi_i(s_i)\phi_j(s_j)\prod_{k\in i\setminus j} m_{ki}(s_i)\prod_{k\in j\setminus i} m_{kj}(s_j). \tag{1.37}$$

As the field of pattern recognition does not use idea of energy $H(\{s_i\})$ we will define it through the Boltzmann distribution for our convenience:

$$P(\{s_i\}) = \frac{1}{Z}e^{-\beta H(\{s_i\})}. \tag{1.38}$$

The Belief Propagation finds the actual probability distribution $P(\{s_i\})$ in an iterative way through the believes $b(\{s_i\})$. It is then helpful to define

some distance measure that would specify how far we are from the actual probability distribution. We will use the Kullback–Leibler distance

$$d\left(b(\{s_i\}), P(\{s_i\})\right) = \sum_{\{s_i\}} b(\{s_i\}) \ln \frac{b(\{s_i\})}{P(\{s_i\})}. \tag{1.39}$$

The measure of distance is non-negative and it is zero only if $b = P$. Substituting the Boltzmann distribution in distance we get

$$d(b, P) = \sum_{\{s_i\}} b(\{s_i\}) \ln b(\{s_i\}) + \beta \sum_{\{s_i\}} b(\{s_i\}) H(\{s_i\}) + \ln Z. \tag{1.40}$$

We know that once we reach the correct probability distribution, the distance becomes zero and thus we have (after the convergence)

$$-\ln Z = \sum_{\{s_i\}} b(\{s_i\}) \ln b(\{s_i\}) + \beta \sum_{\{s_i\}} b(\{s_i\}) H(\{s_i\}), \tag{1.41}$$

what can be immediately rewritten in the form of the second law of thermodynamic $F = E - \frac{1}{\beta} S$, with customary definition of the free energy $F = \ln Z/\beta$.

$$F = \sum_{\{s_i\}} b(\{s_i\}) H(\{s_i\}) - \frac{1}{\beta} \sum_{\{s_i\}} b(\{s_i\}) \ln b(\{s_i\}). \tag{1.42}$$

Bethe Approximation to the Free Energy

Because of the pair-wise form of the probability (1.29) the believes b_i and the two–point believes b_{ij} are sufficient to determine the average energy

$$E = \sum_{\{s_i\}} P(\{s_i\}) H(\{s_i\}), \tag{1.43}$$

which may be rewritten by applying the definition of probability (1.29) and of believes (1.30, 1.37) in the form

$$E = -\sum_{(ij)} b_{ij}(s_i, s_j) \ln \psi_{ij}(s_i, s_j) - \sum_i b_i(s_i) \ln \phi_i(s_i). \tag{1.44}$$

If the graph is tree–like, we can write the joint probability distribution, or the joint belief, in terms of believes b_i and b_{ij} (Bethe approximation)

$$P(\{s_i\}) = \frac{\prod_{(ij)} b_{ij}(s_i, s_j)}{\prod_i b_i(s_i)^{k_i - 1}}, \tag{1.45}$$

where k_i is the number of j's interacting with the property i. In the case of the spin glass it is the number of spins j interacting with the spin i. Using this formula we obtain Bethe approximation to the entropy

$$S = -\sum_{(ij)} \sum_{s_i, s_j} b_{ij}(s_i, s_j) \ln b_{ij}(s_i, s_j) + \sum_i (k_i - 1) \sum_{s_i} b_i(s_i) \ln b_i(s_i). \tag{1.46}$$

The Bethe approximation to the entropy and hence to the free energy is exact on any tree–like structure and it is also a pretty good approximation for the actual entropy on loopy structures [61, 62]. Still, using the standard methods of statical physics (like cluster variational method) we can further improve the approximation.

For our analysis it is important to note the similarity of the equations (1.26) and (1.30) hence of the messages in the Belief Propagation and the cavity biases for the spin glass (note that for our simple spin glass $\phi(s_i) = 1$). We can interpret the calculations done for the spin glass within the replica–symmetric approximation to be equivalent to the Belief Propagation. In the case of the replica–symmetry broken phase we have to develop another method that would take into account the existence of several pure states.

General Survey Propagation

As we have mentioned already, the replica symmetric cavity method may be extended also to problems where the clustering of states occurs. Although we cannot expect that the correlations of distant spins vanish generally, we know that they disappear within each pure state. Hence we can do the calculation similar to the Belief Propagation within each pure state and then summarize somehow the results in all states—do a *survey*—and use the results of the survey for the next iteration. Thus we will move from the probability distribution of cavity fields and biases to surveys of those probability distributions and, in the case of existence of many pure states, to some functional distribution over probability distributions. In the following calculations we will restrict ourselves to the zero–temperature statistics, only.

If we restrict ourselves to a single cluster we may write for the energy of the graph with cavity at zero temperature as before (1.15):

$$E_\alpha^{N-1} = E_{0\alpha}^i \sum_{m\in i} \min_{s_m=\pm 1} h_m^\alpha s_m, \tag{1.47}$$

where we have introduced the cluster index α. For the complete graph we have (as in (1.16))

$$E_\alpha^N = E_{0\alpha}^i + \sum_{m\in i} \min_{s_m=\pm 1} \left\{h_m^\alpha s_m + J_{mi} s_m s_i\right\}, \tag{1.48}$$

which can be expressed in the terms of cavity biases. The only difference is that now the biases will depend on the pure state under consideration:

$$E_\alpha^N = E_{0\alpha}^i + \sum_{m\in i} \left[-w_{mi}^\alpha - u_{mi}^\alpha s_i\right]. \tag{1.49}$$

The local field acting on the spin s_i can be computed (within the pure state) as

$$h_i^\alpha(s_i) = \sum_{m\in i} u_{mi}^\alpha. \tag{1.50}$$

In each state we can evaluate the change of the total energy after filling in the cavity

$$\begin{aligned}\Delta E_\alpha(s_i) &= E_\alpha^N(s_i) - E_\alpha^{N-1} \\ &= \sum_{m\in i}\left[-w_{mi}^\alpha - u_{mi}^\alpha s_i - |h_m^\alpha|\right] \\ &= \sum_{m\in i}\left[-w_{mi}^\alpha - u_{mi}^\alpha s_i - |\sum_{n\in m} u_{nm}^\alpha|\right].\end{aligned} \tag{1.51}$$

In contrast to the Belief Propagation equations (1.28), in the clustered phase we expect that not the probability distributions of h's but the functional measure on $P(h)$ is equal for all the sites. Hence we have to use information from all states in order to find out the functional measure. Still, we have to be careful, since in various states the change of energy ΔE^α differs.

In order to take this under consideration, we will evaluate the probability of finding field h at a site i within the ensemble with a given value of energy. We will consider only final states with some given energy E, yet the number of clusters at varying energies $E - \Delta E_\alpha$ may differ. This we have to consider in our calculation. Let us define quantity called *complexity* Σ that will count the number of pure states with some energy E

$$\text{number of pure states} = e^{N\Sigma(E)}. \tag{1.52}$$

The difference of energetic changes leads to a new term in the equation for $P(h)$, which reflects the level crossing.

To show that, we will calculate the probability distribution $P_i^e(h)$ of finding a field h at the cavity site i with restriction to configurations with some energy density $e = E/N$. We have certainly to count the number of states at a given energy and to weight the contribution with the number of states (at energy $E - \Delta E^\alpha$).

Up to a multiplicative constant we can do it by placing the reweighing term

$$e^{-\Delta[\Sigma(e)]} \simeq e^{-y\Delta E_\alpha} \tag{1.53}$$

in the formula (1.28). We have introduced *pseudo–temperature* y, the temperature–like quantity conjugated to the complexity:

$$y = \frac{\partial\Sigma(e)}{\partial e}. \tag{1.54}$$

Now we can express the probability $P^e(h)$ with the help of formulae (1.51, 1.28)

$$P_i^e(h) = C\int\prod_{m\in i}[P_m(h_m)\mathrm{d}h_m]\,\delta(h_i - \sum_{m\in i}u_{mi})\exp\left\{y\sum_{m\in i}[w_{mi}+h_m] + y|\sum_{m\in i}u_{mi}|\right\}. \tag{1.55}$$

As in the case of the Belief Propagation it may be worth splitting the convolution into two steps, having distributions of the fields P and the biases Q.

$$Q^e(u_{mi}) = \int \mathrm{d}h P_j^e(h)\delta(u_{mi} - u_{mi}(J,h)) \tag{1.56}$$

$$P_i^e(h_i) = \int \prod_{m\in i} [Q^e(u_{mi})\mathrm{d}u_{mi}] \times \exp\{y\sum_{m\in i}[w_{mi}+h_m] + y|\sum_{m\in i} u_{mi}|\}\delta(h_i - \sum_{m\in i} u_{mi}).$$

There are certainly several possibilities how to distribute the reweighing term between the two equations. Our choice is convenient for technical reasons.

We have to keep in mind that P^e is a probability function randomly drawn from some functional distribution $\mathcal{P}(P^e)$ that we are looking for. The same is true for the distributions Q^e. One could certainly write complicated functional equations for the measures $\mathcal{P}$ and $\mathcal{Q}$, and that is actually the path followed by the replica method. We will, on the other side, prefer to find some algorithmic way for finding the distributions of h's and u's. This is done by the Population Dynamics.

The *Population Dynamics* algorithm proceeds as follows:

0) Generate the initial ensemble of probability distributions $\{Q_m^e\}_{m\in i}$ for all sites i.

1) Calculate the probability distributions $P_i^e(h_i)$ according to (1.56).

2) Evaluate Q's following (1.56) and hence obtain some new probability distribution $Q_0(u_{mi})$.

3) Replace one of the initial distributions by the obtained distribution $Q_0(u_{mi})$.

We repeat this procedure until we reach a fixed point and the Q–distributions will not vary virtually. With the fixed–point distributions we can evaluate all the fields h and finally any quantity of our interest.

Still we have to select the appropriate value of pseudo–temperature somehow. What we certainly want to do, is to consider as large number of clusters as possible, and thus to have as representative ensemble as possible. The corresponding choice of energy is then the value at which the complexity has its maximum. As we are looking for the value of the pseudo–temperature y it will be the point where we obtain the maximum of Legendre transform of the complexity. Let us call this transform the 'free energy' $\Phi(y)$.

To calculate the free energy we will follow the consequences of addition of a spin variable, together with all corresponding constraints, to the factor graph. After addition the number of constraint increases by the degree of the spin k, the number of spins increases by one, and the energy increases by ΔE,

see (1.51). Once knowing the probability distributions for the fields we can from this formula also derive what is the probability distribution $P_i(\Delta E)$ of the energetic change,

$$P_i(\Delta E) = \int \prod_{m\in i}[P_i(h_i)\mathrm{d}h_i]\delta\left(\Delta E - \sum_{m\in i}[-w_{mi} - |h_m|] - |\sum_{n\in m} u_{nm}|\right). \quad (1.57)$$

As we know the distribution of possible energetic changes during addition of the spin, we can express the number of states at the end of the manipulation,

$$\exp\left\{N\Sigma(\frac{M}{N},\frac{E}{N})\right\} = \int [P_i(\Delta E)\mathrm{d}(\Delta E)]\exp\left\{(N-1)\Sigma(\frac{M-k}{N-1},\frac{E-\Delta E}{N-1})\right\}, \quad (1.58)$$

which, after expansion of the complexity in $\alpha = \frac{M}{N}$ and $\epsilon = \frac{E}{N}$, leads to

$$\Sigma(\alpha,\epsilon) - \epsilon\frac{\partial\Sigma}{\partial\epsilon} + 2\alpha\frac{\partial\Sigma}{\partial\alpha} = \left\langle \ln\left\{\int [P_i(\Delta E)\mathrm{d}(\Delta E)]e^{-y\Delta E}\right\}\right\rangle_{graphs}. \quad (1.59)$$

We have done averaging over all possible values of k, of the degree of the added node. In the random graph of the Erdős–Renyi kind it is $\langle k\rangle = 2\alpha$.

We see that if we knew the term $\frac{\partial\Sigma}{\partial\alpha}$ we could interpret the previous equation as definition of the Legendre transform of the complexity,

$$\Sigma(\alpha,\epsilon) - y\epsilon = -y\Phi(y), \quad (1.60)$$

with $y = \frac{\partial\Sigma}{\partial\epsilon}$.

The derivative of the complexity with respect to the ratio of constraints α can be calculated by evaluating the change of complexity upon addition of a new constraint to the graph. The energetic change after the addition of the link between spins s_i and s_j is

$$\Delta E_l(h_i,h_j,J_{ij}) = \min_{\{s_i,s_j\}}\{h_i s_i + h_j s_j - J_{ij}s_i s_j\} - |h_i| - |h_j|, \quad (1.61)$$

and hence the induced probability distribution $P_l(\Delta E)$ is

$$P_l(\Delta E_l) = \int [P_i(h_i)P_j(h_j)\mathrm{d}(h_i)\mathrm{d}(h_j)]\delta(\Delta E_l - \Delta E_l(h_i,h_j,J_{ij})). \quad (1.62)$$

Having this distribution we can write for the change of the number of states

$$\exp\left\{N\Sigma(\frac{M+1}{N},\frac{E}{N})\right\} = \int [P_l(\Delta E_l)\mathrm{d}(\Delta E_l)]\exp\left\{N\Sigma(\frac{M}{N},\frac{E-\Delta E_l}{N})\right\}, \quad (1.63)$$

which after expansion of complexity in α on the left–hand side and in ϵ on the right–hand side, and after averaging over possible placement of the link (over graph structures) leads to

$$\frac{\partial \Sigma}{\partial \alpha} = \left\langle \ln \left\{ \int [P_l(\Delta E_l) \mathrm{d}(\Delta E_l)] e^{-y \Delta E_l} \right\} \right\rangle_{graphs}. \tag{1.64}$$

Finally we can define the free energy $\Phi(y)$ as the Legendre transform of the complexity and to express it as $\Phi(y) = \Phi_l^f(y) - (2\alpha - 1)\Phi_i^v(y)$, where the free energy $\Phi(y)$ is separated into the parts corresponding to single variables and single links (factors). The prefactor $2\alpha - 1$ comes from the sequential building of the graph. If we add the factor l to the graph we add also all its variables. And since the mean connectivity in the random graph is 2α we have to subtract the variable contribution exactly $(2\alpha - 1)-$times in order to have it in the correct proportion. The partial terms then read

$$\begin{aligned} \Phi_l^f(y) &= -\frac{1}{y} \left\langle \ln \left\{ \int [P_l(\Delta E_l) \mathrm{d}(\Delta E_l)] e^{-y \Delta E_l} \right\} \right\rangle_{graphs} \\ \Phi_i^v(y) &= -\frac{1}{y} \left\langle \ln \left\{ \int [P_i(\Delta E) \mathrm{d}(\Delta E)] e^{-y \Delta E} \right\} \right\rangle_{graphs}. \end{aligned} \tag{1.65}$$

The optimal value of the pseudo–temperature can be found as the point at which the value of the free energy is maximized.

It is worth noting that the formulae (1.65) are model–independent and that they are valid for any spin glass Hamiltonian that can be represented as a factor graph. Certainly, one has to evaluate the quantities ΔE and ΔE_l in accordance with the model in consideration.

Single Instance Analysis

In our previous calculations we were mainly interested in typical properties of a glassy system. That is, certainly, what we want to do when we are looking for the macroscopical behaviour of a physical system, but in the case of combinatorial optimization we definitely prefer to find complete available information about a single instance. To do so, we have to leave description on the level of statistical ensemble of graphs.

Fortunately enough, the formulae (1.56) may be reconsidered also as formulae applied to a single instance, disregarding the underlying structure averaging. For large enough graphs with a tree-like structure we can still accept validity of the theory.

1.3.4 Heuristic Algorithms

Both the Belief Propagation and the Survey Propagation are in a sense global procedures—the messages are evaluated through the set of update formulae

which includes all the nodes and factor nodes of the graph. On the other side there have been in use many local search algorithms that solve or approximately solve the problems that we are going to work with. It is hence worthy to mention those algorithms which are widely used and with which we will compare the results of the message–passing algorithms.

Monte Carlo & Simulated Annealing

Having some Hamiltonian function $H(\{s_i\})$ that depends on a set of spin variables s_i we can model path of the system in the configuration space by a simulation that simply changes values of variables s_i, with an appropriate acceptance ratio. As we suppose that our system is in thermal equilibrium, we expect that it obeys the *detailed balance* and that the overall distribution of energies will be the Boltzmann distribution of the canonical ensemble at some temperature β. We have to define acceptance criterion in order to meet the two conditions. First we observe that the detailed balance may be met only in the case when the spin flip is reversible. And it is. The acceptance is then often done in the Simulated Annealing (or Monte Carlo) studies through the *Metropolis algorithm*, which goes as follow:

1. Initialization: Generate random configuration of the spins s_i^{init} and calculate its energy $E = H(\{s_i^{init}\})$.

2. Flip: Choose a spin at random and flip its value. Evaluate the change of energy $\Delta E = E^{new} - E$.

3. Decision: If the new energy is smaller, accept the flip and make $E = E^{new}$. Otherwise throw a random number d from range between 0 and 1. If $\exp -\beta\Delta E > d$ accept the flip and the new energy. Otherwise reject the spin flip and return to the configuration and energy before the flip.

The latter two steps of the procedures are then repeated many times in order to sample the configuration space in a representative way. The standard strategy to find what is the minimal number of accepted moves is to evaluate autocorrelation functions of some quantity (energy, magnetization), and to do much more spin flips than what would correspond to the autocorrelation time. For more details see [4, 5].

For the zero temperature and hence $\beta \to \infty$ the algorithm will find the closest local minima of energy and it will get stuck in it. The introduction of temperature and hence of the moves that may increase the total energy enables escaping from a local minimum. The energetic barrier that we can overcome depends on the temperature—the higher temperature we use the larger part of configuration space we sample. On the other side, high temperature will presumably lead to configurations with higher energy, and hence it will prevent us of finding the global minima of the energy landscape.

The technique of Simulated Annealing is used to cross high barriers and on the other side reach low energies. It is usual to introduce a cooling schedule in the algorithm—the temperature decreases after a number of steps in a predefined way. The most efficient schedules are just geometrical series of temperatures where the initial high temperature is decreased by a constant factor after some number of accepted spin flips. For the systems without configuration space clustering and for the system with full-RSB this procedure is guaranteed to reach the global minima for long enough schedules (unfortunately, long enough means doing infinitesimal small changes in temperature). We will see in the case of MAX-K-SAT problem that Simulated Annealing is really efficient in the case of RS and full-RSB phase.

WalkSat

The success of Monte Carlo algorithm and of Simulated annealing in many fields of physics and life sciences was striking. And it is allowance of steps that may increase the total energy, and hence they decrease greediness of the algorithm, that this heuristics is so efficient.

Still, in non-physical applications we are not that much interested in obeying the constraints of detailed balance and of Boltzmann law, there is just no need to follow them. Similarly the choice of the flipped spin may be done in a more specific way when we deal with a problem that has some intrinsic structure.

The *WalkSat* algorithm of Selman, Kautz and Cohen [54] is a heuristics which is strongly influenced by the Monte Carlo heuristic, still it introduces valuable changes to the algorithm that reflects the problem under consideration, the MAX-K-SAT problem.

The WalkSat algorithm mixtures greedy moves with random walk in the configuration space. The algorithm proceeds as follows:

1. Initialization. Create random configuration of spins and calculate the energy of the system.

2. Randomly alternate following steps:

   - Greedy step. Among all variables chose the one that after flipping makes the largest decrease of the energy. Flip it. Update energy.
   - Random walk step. Choose one unsatisfied clause at random and after choose one variable within it. Flip the value of that variable. Recalculate energy. This step may increase the energy actually.

WalkSat stops if either a satisfactory assignment is found or if the maximum number of allowed spin flips (the 'cut-off') is reached.

Chapter 2

The Satisfiability Problem

In the first part of my theses the Survey Propagation algorithm was applied to the combinatorial optimization problem called Max-K-SAT. The collaborative work with Demian Battaglia and Riccardo Zecchina has been published in [6, 7].

2.1 Combinatorial–Optimization Problems and Spin Glasses

In computer science the field of combinatorial optimization [15] deals with the general issue of classifying the computational difficulty ("hardness") of minimization problems and of designing search algorithms. In similarity to statistical physics models, a generic combinatorial optimization problem is composed of many discrete variables—*e.g.*, Boolean variables, finite sets of colours or Ising spins—which interact through constraints typically involving a small number of variables, that in turn sum up to give the global cost-energy function.

As we have mentioned above, among the tools for numerical investigations of complex systems at low temperatures the simulated annealing (SA) algorithm [14] and its variants have played the major role. Such stochastic processes satisfy detailed balance and their behaviour can be compared with static and dynamical mean-field calculations. However, in problems in which the interest is focused on zero temperature ground states and where the proliferation of metastable states causes an exponential slowdown in the equilibration rate, the applicability of SA-like algorithms is limited to relatively small system sizes.

When the problem instances are extracted at random from nontrivial ensembles (that is ensembles which contains many instances that are hard to solve), computer science meets physics in a very direct way: many of the models considered to be of basic interest for Computer Science are nothing but spin glasses defined over finite connectivity random graphs, the well

studied diluted spin glasses [16, 17]. Their associated energy function counts the number of violated constraints in the original combinatorial problem (with ground states corresponding to optimal solutions). Understanding the onset of hardness of such systems is at the same time central to computer science and to $T = 0$ statistical physics with surprisingly concrete engineering applications. For instance, among the most effective error correcting codes and data compression methods are the Low Density Parity Check algorithms [19, 20, 21], which indeed implement an energy minimization of a spin glass energy defined over a sparse random graph. In such problems, the choice of the graph ensemble is a part of the designing techniques, a fact that makes spin glass theory directly applicable.

The above example is however far from representing the general scenario for combinatorial problems: in many situations the probabilistic set up is not defined and, consequently, the notion of typical-case analysis does not play any obvious role. The study of the connection (if any) between worst-case and typical-case complexity is indeed an open one and very few general results are known [22]. Still, a precise understanding of non-trivial random problem instances promises to be important under many aspects. New algorithmic results as well as many mathematical issues have been put forward by the statistical physics studies, with examples ranging from phase transitions [23, 24] and out-of-equilibrium analysis of randomized algorithms [25] to new classes of message-passing algorithms [26, 28].

The physical scenario for the diluted spin glasses version of hard combinatorial problems predicts a trapping in metastable states for exponentially long times of local search dynamic process satisfying detailed balance. Depending on the models and on the details of the process—*e.g.*, cooling rate for SA—the long time dynamics is dominated by different types of metastable states at different temperatures [29]. A common feature is that at zero temperature and for simulation times which are sub-exponential in the size of the problem there exists an extensive gap in energy which separates the blocking states from true ground states.

Such behaviour can be tested on concrete random instances which therefore constitute a computational benchmark for more general algorithms. Of particular interest for computer science are randomized search processes which do not properly satisfy detailed balance and that are known (numerically) to be more efficient than SA-like algorithms in the search for ground states [30]. Whether the physical blocking scenario applies also to these artificial processes, which are not necessarily characterized by a proper Boltzmann distribution at long times, is a difficult open problem. The available numerical results and some approximate analytical calculations [31, 32] seem to support the existence of a thermodynamical gap, a fact which is of up-most importance for optimization. For this reason (and independently from physics), during the last decade the problem of finding the minimal energy configurations of random combinatorial problems similar to diluted spin-glasses—*e.g.*,

random K-Satisfiability (K-SAT) or Graph Coloring—has become a very popular algorithmic benchmark in computer science [17].

In the last few years there has been a great progress in the study of spin glasses over random graphs which has shed new light on mean-field theory and has produced new algorithmic tools for the study of low energy states in large single problem instances. Quite surprisingly, problems which were considered to be algorithmically hard for local search algorithms, like for instance random K-SAT close to a phase boundary, turned out to be efficiently solved by the Survey Propagation (SP) algorithm arising from the replica symmetry broken (RSB) cavity approach to diluted spin glasses.

2.1.1 Definition of the Max-K-SAT Problem

K-SAT is a NP-complete problem [34] (for $K > 2$) which lies at the root of combinatorial optimization. It is very easy to state: Given N Boolean variables and M constraints taking the form of clauses, *K-SAT* consists in asking whether it exists an assignment of the variables that satisfies all constraints. Each clause contains exactly K variables, either directed or negated, and its truth value is given by the OR function. Since the same variable may appear directed or negated in different clauses, competitive interactions among clauses may set in.

As mentioned in the introduction, in the last decade there has been a lot of interest on the random version of K-SAT: for each clause the variables are chosen uniformly at random (with no repetitions) and negated with probability $\frac{1}{2}$. The random-MAX-K-SAT problem is then the optimization problem derived from the random-K-SAT. The task is to find an assignment to all variables that would minimize the total number of violated clauses in the formula.

2.1.2 The Phase Diagram of the Random Max-K-SAT Problem

In the large N limit, random K-SAT displays a very interesting threshold phenomenon. Taking as a control parameter the ratio of number of clauses to number of variables, $\alpha = M/N$, there exists a phase transition at a finite value $\alpha_c(K)$ of this ratio. For $\alpha < \alpha_c(K)$ the generic problem is satisfiable (SAT), for $\alpha > \alpha_c(K)$ the generic problem is not satisfiable (UNSAT).

This phase transition has been seen numerically [35] and it is of special interest since extensive experiments [17] have shown that the instances, which are algorithmically hard to solve, are exactly those where α is close to α_c. Therefore, the study of the SAT/UNSAT phase transition is considered of crucial relevance for understanding the onset of computational complexity in typical instances [16]. A lot of work has been focused on the study of both

the decision problem (*i.e.*, determining with a YES/NO answer whether a satisfying assignment exists), and the optimization version in which one is interested in minimizing the number of violated clauses when the problem is UNSAT (random MAX-K-SAT problem).

On the analytical side, there exists a proof that the threshold phenomenon exists at large N [36], although the fact that the corresponding α_c has a limit when $N \to \infty$ has not yet been established rigourously. Upper bounds $\alpha_{\rm UB}(K)$ on α_c have been found using first moment methods [37] and variational interpolation methods [38], and lower bounds $\alpha_{\rm LB}(K)$ have been found using either explicit analysis of some algorithms [39], or some second moment methods [40]. For random MAX-K-SAT theoretical bounds are also known [41, 45], as well as rigourous results on the running times of random walk and approximation algorithms [42, 43, 44].

Recently, the cavity method of statistical physics has been applied to K-SAT [24, 26, 46, 7] and the thresholds have been computed with high accuracy. A lot of work is going on in order to provide a rigourous foundation to the cavity results and we refer to [46] for a more complete discussion of these aspects.

In what follows we shall concentrate on the $K = 3$ case and we will be interested in analyzing the behaviour of different algorithms in the region of parameter α in which the random formulae are expected to be hard to solve or to minimize. The energy function which is used in the zero temperature statistical mechanics studies is taken proportional to the number of violated clauses in a given problem so that a zero energy ground state corresponds to a satisfying assignment. The energy of a single clause is positive (equals 2 for technical reasons) if the clause is violated and zero if it is satisfied. The overall energy is obtained by summing over clauses and reads

$$E = 2 \sum_a \frac{\prod_{i=1}^3 (1 + J_{a,i} s_i^a)}{2} \tag{2.1}$$

where s_i^a is the i-th binary (spin) variable appearing in clause a and the coupling $J_{a,i}$ takes the value 1 (resp. -1) if the corresponding variable appears not negated (resp. negated) in clause a. For instance the clause $(x_1 \vee \bar{x}_2 \vee x_3)$ has an energy contribution $\frac{1}{4}(1 + s_1)(1 - s_2)(1 + s_3)$ where the Boolean variables $x_i = \{0, 1\}$ are connected to the spin variables by the transformation $s_i = (-1)^{x_i}$.

2.1.3 Ergodicity Breaking in K-SAT

With increasing number of constraints (clauses) the number of solutions decreases and on the other hand the number of metastable solutions increases. The configuration space also breaks in distinct states.

The phase diagram of the random 3-SAT problem as arising from the statistical physics studies can be very briefly summarized as follows.

For $\alpha < 3.86$, the $T = 0$ phase is at zero energy (the problem is SAT). The entropy density is finite and the phase is Replica Symmetric (RS) and unfrozen. Roughly speaking, this means that there exists one giant cluster of nearby solutions and that the effective fields vanish linearly with the temperature.

For $3.86 < \alpha < 3.92$, there is a full RSB phase. The solution space breaks in clusters and the order parameter becomes a nested probability measure in the space of probability distribution describing cluster to cluster fluctuations. The phase is still SAT and unfrozen [47, 48].

At $\alpha \simeq 3.92$ there is a discontinuous transition toward a phase which is currently under discussion. Originally it was considered to be a f-RSB frozen phase [24, 26], but most recent studies are showing that the phase is 1RSB unstable rather for reasons of instability of cavity formulae than due to the instability of 1RSB ansatz (similarly to [49]).

Above $\alpha = 4.15$ the 1-RSB solution of cavity equations becomes stable again [50]. The *complexity*, that is the normalized logarithm of the number of clusters, is finite in this region. At finite energy there exist many metastable states which act as dynamical traps. The 1-RSB metastable states become unstable at some energy density $E_G(\alpha)$ which constitutes a lower bound to the true dynamical *threshold energy* (see Sec. 2.2 for details).

At $\alpha = 4.2667$ the ground state energy becomes positive and therefore the typical random 3-SAT problem becomes UNSAT. At the same point the zero–energy complexity vanishes. The phase remains 1-RSB up to $\alpha = 4.39$ where an instability toward a zero complexity full RSB phase appears.

In the region $4.15 < \alpha < 4.39$, the 1-RSB ansatz for the ground state is stable against higher orders of RSB, but the 1-RSB predictions become unstable for energies larger than the *Gardner energy*. The instability line intersects with the 1-RSB ground state estimation at the two extremes of the interval, inside which it provides a lower bound to the true threshold energy (see Ref. [50] for a comprehensive discussion).

Further (preliminary) f-RSB corrections suggest that the true threshold states have energies very close to the lower bound and hence the interval $A = [4.15, 4.39]$ should be taken as the region where to take really hard benchmarks for algorithm testing.

As displayed in Fig. 2.1, the actual value of the energy gap is very small close to the end points of A. In order to avoid systematic finite size errors, numerical simulations should be done close to the SAT/UNSAT point, *i.e.*, far from the end point of A. Consistently with the fact that finite size fluctuations are relatively big ($O(\sqrt{N})$), even close to α_c problem sizes of the order at least of $N = 10^5$ are necessary in order to observe a matching with the analytic predictions.

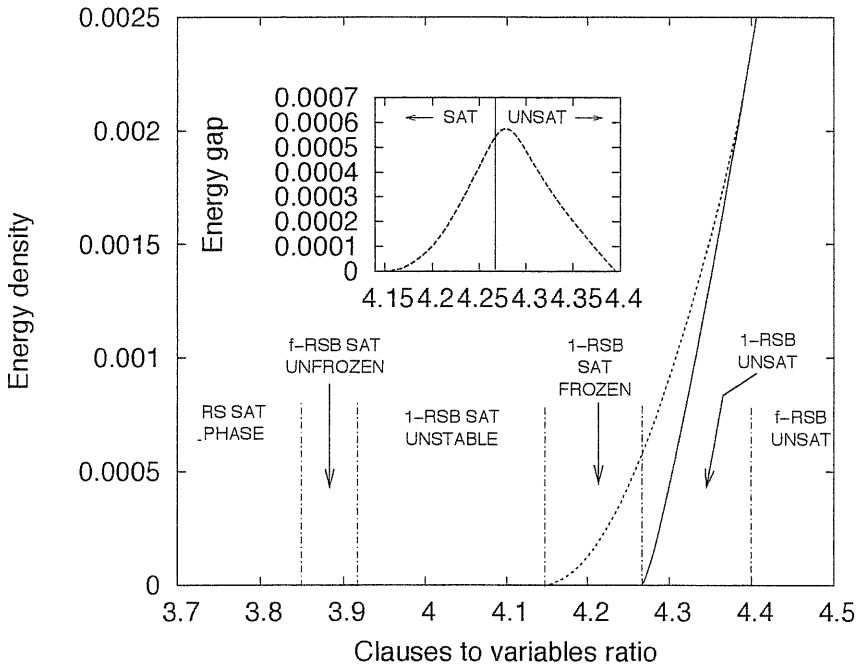


Figure 2.1: The solid line is an estimate for the ground state energy, while the dashed curve represents the Gardner energy, providing a lower bound for the threshold states (numerical data adapted from ref. [50]). In the inset we show that the difference between the Gardner and the ground state energy is strictly positive in the small 1-RSB stable region around the SAT/UNSAT transition critical point (indicated by the vertical line): it is expected that it is hard for heuristics based on local search to find assignments inside the closed area delimited by the energy gap curve.

2.2 The Survey Propagation

2.2.1 The Zero Pseudo-Temperature Survey Propagation—K-SAT

The 1-RSB cavity equations which have been used to study the typical phase diagram of random K-SAT become the SP equations once reformulated to run over single problem instance [26]. This is done by avoiding the averaging process with respect to the underlying random graphs. Thanks to the self-averaging property of the random K-SAT free energy [53], the SP equations can be used both to re-derive the phase diagram of the problem and, more importantly, to access detailed information of algorithmic relevance about a given problem instance. In particular, the SP equations provide information about the statistical behaviour of the single variables in the stable and metastable states of given energy density.

The 1-RSB cavity equations, see Section 1.3.3, are iterative equations (averaged over the disorder) for the probability distribution functions (pdf) of effective fields that describe their cluster-to-cluster fluctuations. The order parameter is a probability measure in the space of pdf's; it tells the probability that a randomly chosen variable has a certain associated pdf in states at a given energy density.

In SP and more in general in the cavity approach, one assumes to know pdf's of the fields of all variables in the temporary absence of one of them. Then one writes the induced pdf of the local field acting on this "cavity" variable in absence of some other variable interacting with it (*i.e.*, the so called Bethe lattice approximation for the problem). These relations define a closed set of equations for the pdf's that can be solved iteratively. The equations are exact if the cavity variables acting as inputs are uncorrelated, *e.g.*, over trees, or are conjectured to be an asymptotically exact approximation over locally tree-like structures [26] where the typical distance between randomly chosen variables diverges in the large N limit (as $\ln N$ for diluted random graphs). The full list of the cavity fields over the entire underlying graph, in the SP implementation, constitutes the order parameter. From the cavity fields one may determine the total field acting on each variable in all metastable states of given energy density and this information can be used for algorithmic purposes.

A clear formalism for the single sample analysis is given by the factor graph representation [27] of K-SAT: variables are represented by N circular "variable nodes" labelled with letters $i, j, k, \ldots$ whereas the K-body interactions are represented by M square "function nodes" (carrying the clause energies) labelled by $a, b, c, \ldots$ (see Fig. 2.2)

For random 3-SAT, function nodes have connectivity 3, variable nodes have a Poisson connectivity of average 3α and the overall graph is bipartite. The total energy is nothing but the sum of energies of all function nodes as

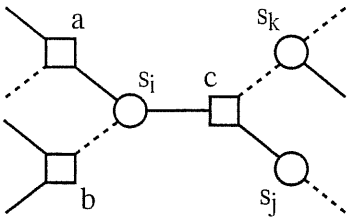


Figure 2.2: Factor graph representation. Variables are represented by circles, and are connected by function nodes, represented by squares; if a variable appears negated in a clause, the connecting line is dashed.

given by Eq. (2.1).

The Cavity Bias for Random 3-SAT

Here we could repeat the analysis of the RS case of the spin glass done in the Introduction in the Section 1.3.3. As the changes are minimal, due to the universality of the formulae for the Believe Propagation, we will summarize only the differences between the two problems.

The factor graph, see Figure 1.6, of the simple spin glass had function nodes with connectivity 2, which has reflected the couplings J_{ij}. In the case of 3-SAT, Figure 2.2, the couplings in the Hamiltonian couple three generally different spins and that is reflected in the structure of the factor graph by function nodes of degree 3. This will cause also difference in the formulae for the cavity biases, as there are two spins neighbouring to a spin i through the function node a. Still the structure of formulae will remain close to what we have seen for simple spin glass with the following changes.

The main differences occur in the expressions for the cavity biases u_{ai} which now has to be indexed by the function node instead of the simple indexing by the neighbouring spin. Similarly, the constants w_{ai} will change in accordance with the Hamiltonian.

$$u_{a\to i}\left(\{h_{j\to a}\}\right) = J_{a,i}\prod_{j\in a\setminus i}\theta\left(J_{a,j}h_{j\to a}\right), \tag{2.2}$$
$$w_{a\to i}\left(\{h_{j\to a}\}\right) = \sum_{j\in a\setminus i}|h_{j\to a}| - \prod_{j\in a\setminus i}\theta\left(J_{a,j}h_{j\to a}\right).$$

Because the basic cavity element in the 3-SAT problem will be a spin variable together with all function nodes to which it belongs, the energetic changes after filling in the cavity or after addition of a function node to the graph will differ. We needed these quantities for finding the formulae for the free energy and for optimization of the pseudo–temperature y in (1.65). The expressions

for the energetic changes may be found in a straightforward way in close similarity to the analysis of the Section 1.3.3. The resulting formulae are

$$\Delta E = \sum_{a \in i} \left[-w_{a \to i} + \sum_{j \in a \setminus i} |h_{j \to a}| \right] - \left| \sum_{a \in i} u_{a \to i} \right|, \tag{2.3}$$

$$\Delta E_a = \min_{\{s_i\}, i \in a} \left\{ -\sum_{j \in a} h_{j \to a} + 2 \frac{\prod_{j \in a} (1 + J_{a,j} s_j)}{2} - \sum_{j \in a} |h_{j \to a}| \right\},$$

where the only difference is that we index the change after adding a function node by index of the node rather than by the link index which in the 3-SAT case looses meaning. For detailed analysis see [26].

All the other formulae are universal as far as the Hamiltonian can be represented by a factor graph.

Adopting the message-passing notation and strictly following [26], we call u-messages the contribution to the cavity fields coming from the different connected branches of the graph. In SP the messages along the links of the factor graph have a functional nature carrying information about distributions of u-messages over the states at a given value of the energy, fixed by a Lagrange multiplier y: we call these distributions of messages u-surveys. The SP equations can be written at any "temperature" (the inverse of the Lagrange multiplier y is actually a pseudo-temperature, see [26]). However they acquire a particularly simple form in the limit $1/y \to 0$, which is the limit of interest for optimization purposes, at least in the SAT region.

In K-SAT, the u-surveys are parameterized by two real numbers and SP can be implemented very efficiently. Each edge $a \to i$, from a function node a to a variable node i, carries a u-survey $Q_{a \to i}(u)$. From these u-surveys one can compute the cavity fields $h_{i \to b}$ for every $b \in i$, which in turn determine new output u-surveys (see Fig. 2.3).

Very schematically, the SP equations can be implemented as follows. Given a random initialization of all the u-surveys $Q_{a \to i}(u)$, the function nodes are selected sequentially at random and the u-surveys are updated according to a complete set of coupled functional equations (see Fig. 2.3 for the notation):

$$P_{j \to a}(h_{j \to a}) = C_{j \to a} \int \mathcal{D}Q_{j,a} \; \delta\Big(h - \sum_{b \in j \setminus a} u_{b \to j}\Big)$$

$$\times \exp\Big(y \big(\big|\sum_{b \in j \setminus a} u_{b \to j}\big| - \sum_{b \in j \setminus a} |u_{b \to j}|\big)\Big), \tag{2.4}$$

$$Q_{a \to i}(u) = \int \mathcal{D}P_{a,i} \; \delta\left(u - \hat{u}_{a \to i}\left(\{h_{j \to a}\}\right)\right), \tag{2.5}$$

where the $C_{i \to a}$'s are normalization constants, and the integration measures are given by:

$$\mathcal{D}Q_{j,a} = \prod_{b \in j \setminus a} Q_{b \to j}(u_{b \to j}) \, du_{b \to j}, \tag{2.6}$$

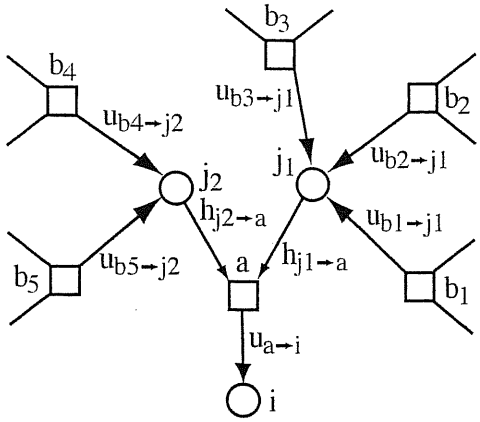


Figure 2.3: Cavity fields and u-messages. The u-survey for the u-message $u_{a\to i}$ depends on the pdf's of the cavity fields $h_{j_1\to a}$ and $h_{j_2\to a}$. These are on the other side dependent on the u-surveys for the u-messages incoming to the variables j_1 and j_2.

$$\mathcal{D}P_{a,i} = \prod_{j\in a\setminus i} P_{j\to a}(h_{j\to a})\, dh_{j\to a}. \tag{2.7}$$

Parameterizing the u-surveys as

$$Q_{a\to i}(u) = \eta^0_{a\to i}\delta(u) + \eta^+_{a\to i}\delta(u-1) + \eta^-_{a\to i}\delta(u+1) \tag{2.8}$$

where $\eta^0_{a\to i} = 1 - \eta^+_{a\to i} - \eta^-_{a\to i}$, the above set of equations (2.4, 2.5) defines a non-linear map over the η's.

Once a fixed point is reached, from the list of the u-surveys one may compute the normalized pdf of the *local field* acting on each variable:

$$P_i(H) = C_i \int \mathcal{D}\widehat{Q}_i\ \delta\Big(H - \sum_{b\in i} u_{b\to i}\Big) \times \exp\Big(y\big(\sum_{b\in i} u_{b\to i}| - \sum_{b\in i}|u_{b\to i}|\big)\Big), \tag{2.9}$$

$$\mathcal{D}\widehat{Q}_i = \prod_{b\in i} Q_{b\to i}(u_{b\to i})\, du_{b\to i}. \tag{2.10}$$

It should be remarked that $P_i(H)$ is in general different from the family of *cavity fields* pdf's $P_{i\to b}(h)$ computed by mean of (2.4).

From the knowledge of the cavity and local fields pdf's, one derives the (Bethe) free energy at the level of 1-RSB:

$$\Phi(y) = \frac{1}{N}\left(\sum_{a=1}^{M}\Phi^f_a(y) - \sum_{i=1}^{N}\Phi^v_i(y)(\Gamma_i - 1)\right), \tag{2.11}$$

where Γ_i is the connectivity of the variable i and:

$$\Phi_a^f(y) = -\frac{1}{y}\ln\Big\{\int\prod_{i\in a}\mathcal{D}Q_{i,a}\ \times \tag{2.12}$$
$$\exp\left[-y\min_{\{\sigma_i,i\in a\}}\left(E_a-\sum_{i\in a}\left[\sum_{b\in i\setminus a}u_{b\to i}\right]\sigma_i+\sum_{b\in i\setminus a}|u_{b\to i}|\right)\right]\Big\},$$
$$\Phi_i^v(y) = -\frac{1}{y}\ln\Big\{\int\mathcal{D}\widehat{Q}_i\ \exp\left[y(|\sum_{a\in i}u_{a\to i}|-\sum_{a\in i}|u_{a\to i}|)\right]\Big\} = -\frac{1}{y}\ln(C_i).$$

Note that (2.11) differs slightly from our definition of free energy in the Introduction. The change reflects that we are considering a single instance and hence we build the graph by adding the variable nodes and the function nodes. When adding the function nodes, Γ_i of them for variable i, we add also the variable several times. This we have to take into account and subtract this fake contributions from the overall free energy. Hence we have the term $\Gamma_i - 1$ in front of the variable contribution. In the case of ensemble of random graphs the equivalent operation was to add the multiplicative factor $2\alpha - 1$. The E_a is the energy contribution of the function node a. The maximum value of the free-energy functional provides a lower bound estimation of the ground state energy of the Hamiltonian (2.1) defined on the sample. In the SAT region the free-energy functional $\Phi(y)$ is always non positive and it is an increasing function of y, the choice $y \to \infty$ is optimal; in the UNSAT region, on the contrary, it exhibits a positive maximum for some $y = y^*$ (see [26]).

From the free-energy density of a given instance, it is straightforward to compute numerically its complexity $\Sigma(y) = \partial\Phi(y)/\partial(1/y)$ and its energy density $\epsilon(y) = \partial(y\Phi(y))/\partial y$, the two quantities are connected through the Legendre transform (1.60). We remind that the complexity is linked to the number of pure states (*i.e.*, clusters of configurations) of energy E, by the defining relation $\mathcal{N}(E) = \exp(N\Sigma(E))$. The energy level represented by the largest number of configurations is given by:

$$\Sigma(e_{th}) = \max_E\{\Sigma(E)\}. \tag{2.13}$$

Further RSB corrections may be needed to locate the precise value of e_{th}, which is in any case lower bounded by the largest energy of 1-RSB stable states, the so called *Gardner energy* E_G. It is expected that local search strategies get trapped at energies close, but not necessarily equal, to the threshold energy (see refs. [29] for a thorough discussion on the role of the iso-complexity states [51]). More elaborated strategies not properly satisfying detailed balance (*e.g.*, WalkSat for the K-SAT problem) could in principle overcome this type of barriers; however, the available numerical and analytical

results suggest that also these more sophisticated randomized searches undergo an exponential slowdown, with different layers of states acting as dynamical traps, depending on the details of the heuristics. Although the field is still very active and the results are not well established, see for example [52].

2.2.2 The Finite Pseudo-Temperature Survey Propagation—MAX-K-SAT

In the SAT phase, where the $y \to \infty$ limit is taken, the convolutions (2.4) filter out completely any clause-violating truth value assignment. This feature is extremely useful for satisfiable formulae, but it becomes undesired when our sample is presumably unsatisfiable.

In the UNSAT region the SP equations require a finite value of the Lagrange multiplier y. The filtering action of the exponential re-weighting term in (2.4) is then weakened and the messages computed by the SP equations can vehicle information pointing to states with a non vanishing number of violated constraints.

2.2.3 The Finite Pseudo-Temperature Recursive Equations

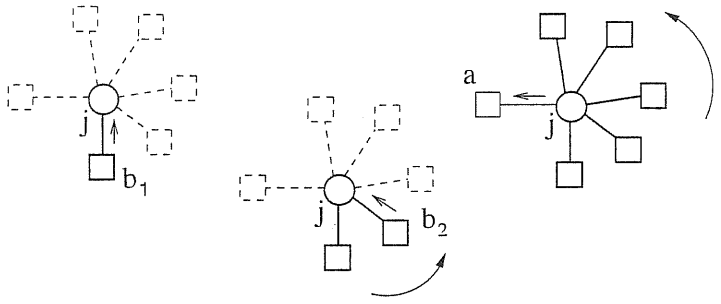
The SP equations simplify considerably in the $y \to \infty$ limit and lead to extremely efficient algorithmic implementations, as discussed in great detail in [28]. In the case of finite pseudo-temperature $1/y$ the same simplification cannot take place because of the presence of a nontrivial re-weighting factor. Still, a relatively fast recursive procedure can be written. Let us consider a variable j having Γ_j neighbouring function nodes and let us compute the cavity field pdf $P_{j\to a}(h)$ where $a \in j$. We start by randomly picking up one function node $b_1 \in j \setminus a$, and we calculate the following "h-survey":

$$P^{(1)}_{j\to a}(h) = \eta^{0}_{b_1\to i}\,\delta(h) + \eta^{+}_{b_1\to i}\,\delta(h-1) + \eta^{-}_{b_1\to i}\,\delta(h+1). \tag{2.14}$$

The function $P^{(1)}_{j\to a}(h)$ would correspond to the true local field pdf of the variable j in the case in which b_1 was the only neighbouring clause (as denoted by the upper index).

The following steps of the recursive procedure consist in adding the contributions of all the other function nodes $b_\gamma \in j \setminus a$, clause by clause (Fig. 2.4):

$$\begin{aligned}\widetilde{P}^{(\gamma)}_{j\to a}(h) = {} & \eta^{0}_{b_\gamma\to j}\,\widetilde{P}^{(\gamma-1)}_{j\to a}(h) \\ & + \eta^{+}_{b_\gamma\to j}\,\widetilde{P}^{(\gamma-1)}_{j\to a}(h-1)\,\exp\left[-2y\,\hat{\theta}(-h)\right] \\ & + \eta^{-}_{b_\gamma\to j}\,\widetilde{P}^{(\gamma-1)}_{j\to a}(h+1)\,\exp\left[-2y\,\hat{\theta}(h)\right].\end{aligned} \tag{2.15}$$



(a) (b) (c)

Figure 2.4: Computing recursively a cavity pdf. (a) In order to find a single cavity pdf $P_{j\to a}(h)$, a single clause $b_1 \in j \setminus a$ is picked up at random and the u-survey $Q_{b_1\to j}$ is used to compute equation (2.14); (b) The contributions of all the other function nodes $b_\gamma \in j \setminus a$ are then added, clause by clause; (c) The pdf computed recursively after $\Gamma_j - 1$ iterations coincides with $P_{j\to a}(h)$.

Here $\widetilde{P}^{(\gamma)}_{j\to a}(h)$ is an unnormalized pdf and $\hat{\theta}(h)$ is a step function equal to 1 for $h \geq 0$ and zero otherwise. The recursion ends after $\gamma = \Gamma_j - 1$ steps, when the influence of every clause $b_\gamma \in j \setminus a$ has been taken in account. The final cavity-field pdf $P_{j\to a}(h)$ can be found straightforwardly by computing the pdf $\widetilde{P}^{(\Gamma_j-1)}_{j\to a}(h)$ for all values of the field $-\Gamma_j + 1 < h < \Gamma_j - 1$ and by normalizing it.

As already pointed out in Section 2.2, the knowledge of $K-1$ input cavity-field pdf's can be used to obtain a single output u-survey. Let us compute for instance the u-survey $Q_{a\to i}(u)$ (see always Fig. 2.3 for the notation). In order to do that, we need first the cavity field pdf's $P_{j\to a}(h)$ for every $j \in a \setminus i$. The parameters $\{\eta^0_{a\to i}, \eta^+_{a\to i}, \eta^-_{a\to i}\}$ are then updated according to the formulae:

$$\eta^{J_{a,i}}_{a\to i} = \prod_{n=1}^{K-1} W^{J_{jn,a}}_{jn\to a}, \quad \eta^{-J_{a,i}}_{a\to i} = 0, \quad \eta^0_{a\to i} = 1 - \eta^{J_{a,i}}_{a\to i}, \tag{2.16}$$

where we introduced the weight factors:

$$W^+_{j\to a} = \sum_{h=1}^{\Gamma_j-1} P_{j\to a}(h), \quad W^-_{j\to a} = \sum_{h=-\Gamma_j+1}^{-1} P_{j\to a}(h). \tag{2.17}$$

It should be remarked that $Q_{a\to i}(u)$ depends only on one single nontrivial $\eta^{J_{a,i}}_{a\to i}$ (from now simply referred to as $\eta_{a\to i}$). We could say that a single kind of message can be produced, telling the receiver literal to assume the truth value "TRUE"; this message is transmitted along the edge $a \to i$ with a probability

$\eta_{a\to i}$, corresponding to the probability that the only way of not violating the constraint a is to set appropriately the truth value of i.

Starting from a full collection of u-surveys at a given time, it is possible to realize a complete update of all the parameters $\{\eta_{a\to i}\}$ by systematical application of the recursions (2.14), (2.15) and of the relation (2.16); from the new set of u-surveys, new cavity field pdf's can be computed and the procedure continues until when self-consistence of η's is reached. This procedure can be efficiently implemented numerically and allows us to determine the fixed point of the population-dynamics equations (2.4), (2.5), for a general value of y.

2.3 Numerical Results

2.3.1 The SP-Y Algorithm

In the usual SP-inspired decimation [28], the computation of the local field pdf's $P_i(H)$ is used to decide a truth value assignment for the most biased variables. Indeed, it is reasonable that a spin tends to align itself with the most probable direction of the local field. A ranking can be realized by finding all the probability weights

$$W_j^+ = \sum_{H=1}^{\Gamma_j} P_j(H), \quad W_j^- = \sum_{H=-\Gamma_j}^{-1} P_j(H), \tag{2.18}$$

and by sorting the variables according to the values of a bias function:

$$b_{\text{fix}}(j) = |W_j^+ - W_j^-|. \tag{2.19}$$

The local field pdf's $P_j(H)$ can be naturally calculated resorting to the iterations (2.14), (2.15): computation is done simply by sweeping over the whole set of neighbouring function nodes $b \in j$, including also the contribution of the skipped edge $a \to j$. By fixing in the right direction the spin of the most biased variable, we actually reduce the original N variable problem to a new one with $N-1$ variables. New u-surveys are then computed. Doing that we have to take care of fixed variables: if i is fixed, its cavity field pdf's must be of the form:

$$P_{i\to a}(h) = \delta\left(h - J_{a,i}s_i\right); \tag{2.20}$$

regardless of the recursions (2.14), (2.15). The complete polarization reflects the knowledge of the truth value of the literals depending on the spin s_i.

The procedure of decimation continues until when a full truth assignment has been generated or until when convergence has been lost or a paramagnetic state has been reached; in the latter cases the original formula is simplified according to the partial truth assignment already generated and the

simplified formula is passed to a specialized heuristic. Our choice of preference is the WalkSat algorithm [54], which is by far more efficient than SA in the hard region of the 3-SAT problem, as we have checked extensively. Very briefly, the strategy of WalkSat is the following one: at each time step the current assignment is changed by randomly alternating greedy moves (where the variable which maximizes the number of satisfied clauses if fixed) and random-walk steps (in which a variable belonging to a randomly chosen unsatisfied clause is selected and flipped). WalkSat stops if either a satisfying assignment is found or if the maximum number of allowed spin flips (the "cutoff") is reached (see Ref. [55] for another recently analyzed and very efficient heuristics).

When working at finite pseudo-temperature, we have to take in account the possibility that some non-optimal fixing is done in presence of 'thermal' noise. After several updates of the u-surveys some biases of fixed spins may become smaller than the value they had at the time when the corresponding spin was fixed. Certain local fields can even revert their orientation. Small or positive values of an index function like:

$$b_{\text{backtrack}}(j) = -s_j \left(W_j^+ - W_j^- \right), \tag{2.21}$$

can track the appearance of such dangerous fixed spins and this information can be used to implement some "error removal" procedure; for instance, a simple strategy can be devised where both unfixing and fixing moves are performed at a fixed ratio $0 \leq r < 0.5$ (see [33] for another backtracking implementation).

The actual SP-Y simplification procedure will depend not only on the backtracking fraction r, but even more on the choice of the pseudo-temperature y. The simplest possibility is to keep it fixed during the simplification, but one may choose to dynamically update it, in order to stay as close as possible to the maximum y^* of the free energy functional $\Phi(y)$.

The equations (2.11, 2.13) can be rewritten in the following form, suitable for numerical computation:

$$\Phi_a^f(y) = -\frac{1}{y}\Big[\ln\Big(1 + (\mathrm{e}^{-y} - 1)\prod_{i\in a} W_{i\to a}^{J_{a,i}}\Big) - \ln\Big(\prod_{i\in a} C_{i\to a}\Big)\Big], \tag{2.22}$$

$$\Phi_i^v(y) = -\frac{1}{y}\ln(C_i)\,. \tag{2.23}$$

In Fig. 2.5 we give a summary of the simplification procedure in a standard pseudo-code notation. The public release of the SP-Y code can be downloaded from [56].

INPUT: a Boolean formula $\mathcal{F}$ in conjunctive normal form; a backtracking ratio r; optionally, a fixed inverse pseudo-temperature y_{in}

OUTPUT: a simplified Boolean formula $\mathcal{F}'$ in conjunctive normal form (ideally empty) and a partial truth value assignment for the variables of $\mathcal{F}$ (ideally a complete one)

0. For each edge $a \longrightarrow i$ of the factor graph, randomly initialize the $\eta_{a\to i} \in \{0,1\}$

1. IF there is a fixed y_{in} as input, put $y^* = y_{in}$, ELSE after a fixed number of steps, determine by bisection the position of the free-energy maximum y^*

2. Compute all the fixed point u-surveys, using equations (2.14), (2.15), (2.16) and putting $y = y^*$

3. IF the population dynamics equations converge,

   3.1 FOR every unfixed variable i, compute the local field pdf using (2.14), (2.15)

   3.2 Extract a random number q in [0,1]

   3.3 IF $q \geq r$, Sort the variables according to the index function (2.19), and fix the most biased variable

   3.4 ELSE IF $q < r$, Sort the variables according to the index function (2.21) and unfix the highest ranked variable

   3.5 IF all the variables are fixed, RETURN the full truth value assignment and an empty sub-formula, ELSE, go to 1.

4. ELSE IF the population dynamics equations do not converge, simplify the formula by imposing the already assigned truth values, RETURN the partial solution and the obtained sub-formula

Figure 2.5: The SP-Y simplification algorithm.

2.4 Optimizing Energy Below the Threshold States

As we have already discussed in Section 2.2, it is expected that, in the thermodynamical limit, any local search algorithm gets trapped in the vicinity of exponentially numerous threshold states with energy e_{th} and that any local heuristics is in general unable to find the optimal assignment in the thermodynamical limit. To verify this prediction, we conducted various experiments, both in the SAT and in the UNSAT phase, focusing on the comparison between the WalkSat heuristics performance after and before different kinds of SP-Y simplification. In most of the situations, we decided to analyze carefully single large-sized samples instead of a larger number of smaller problems: we verified in fact that the sample-to-sample fluctuations tend to be irrelevant for size of order 10^4 and larger.

2.4.1 SAT Region

The aim of the first set of experiments was to check the actual existence of the threshold effect. We ran WalkSat over different formulae in the hard-SAT region, with fixed $\alpha = 4.24$ and sizes varying between $N = 10^3$ and $N = 10^5$, reaching a maximum cut-off of 10^{10} spin flips. The obtained results are plotted in Fig. 2.6; the Gardner energy is also reported for comparison with the data. Even if for small-size samples the local search algorithm is able to find a SAT assignment, for larger formulae ($N \sim \mathcal{O}(10^4)$) WalkSat does not succeed in reaching the ground state, its relaxation profile suffers of critical slowdown, and saturates at some well defined level. This is actually expected, because the Gardner energy becomes $\mathcal{O}(1)$ only for $N \sim 10^4$ or larger, and for a smaller number of variables the threshold effect should be negligible when compared to finite size effects.

We remind that WalkSat cannot be considered as an equilibrium stochastic process and that it is not possible to infer that its saturation level coincides with the sample threshold energy; we can anyway claim that WalkSat is unable to explore the full energy landscape of the problem, and that the enormous number of non optimal valleys is unavoidably hiding the true ground states. Plateaux in the relaxation profiles of WalkSat have indeed been already discussed in [31, 32] and ascribed to metastable states acting as dynamical traps.

For the $N = 10^4$ formula a trapping effect becomes clearly visible in our experiments, but the saturation plateau is below the Gardner lower bound. The finite-size fluctuations are still of the same order of the energy gap between the ground and the threshold states and the experimental conditions are distant from the thermodynamical limit. When the size is increased up to 10^5 variables, the saturation level moves finally between the full RSB lower bound and the 1-RSB upper bound for e_{th}.

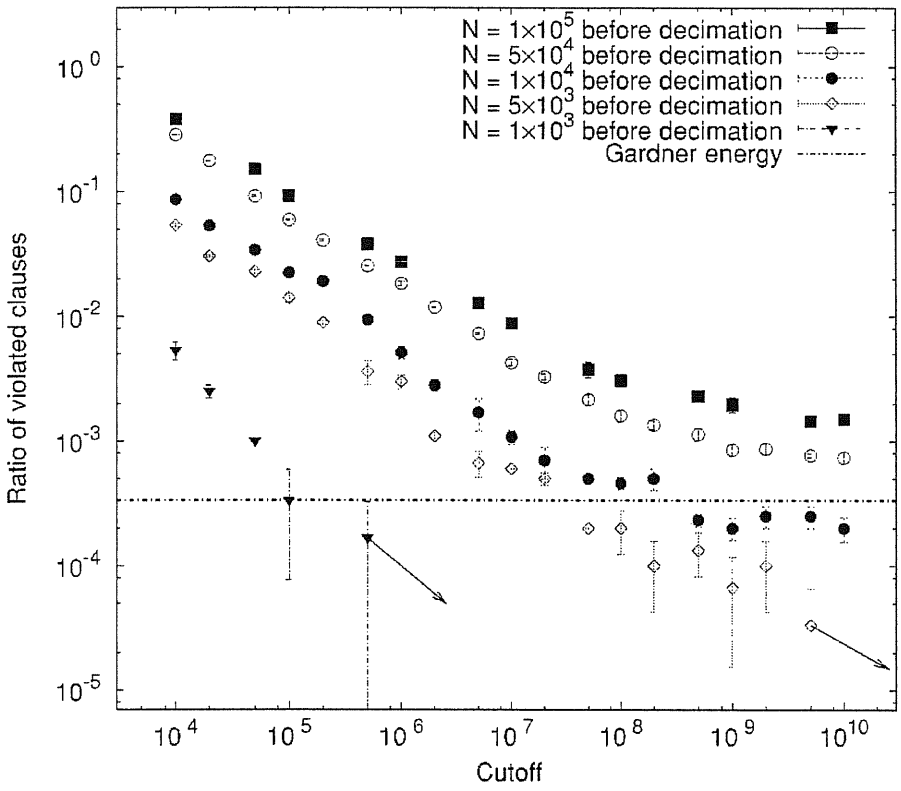


Figure 2.6: Threshold energy effect in SAT region. The WalkSat performance for various samples of different sizes and $\alpha = 4.24$ is presented. With increasing size, the curves appear to saturate above the Gardner energy. An arrow indicates that the next data point corresponds to a SAT assignment.

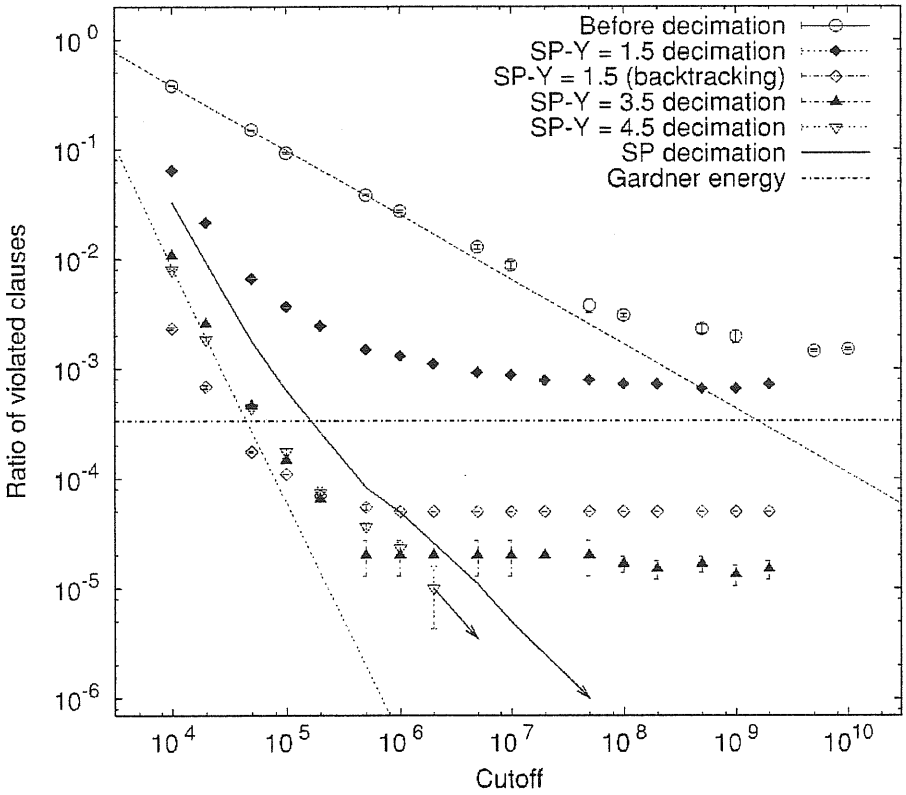


Figure 2.7: Efficiency of SP-Y in the SAT region (single sample with $N = 10^5$ variables and $\alpha = 4.24$). After SP-Y simplification, WalkSat is generally able to find solutions below the Gardner threshold; in some cases, it succeeds even in finding complete satisfying assignment. An arrow indicates that the next data point corresponds to a SAT assignment.

The efficiency of the SP-Y simplification strategy against the glass threshold is discussed in Fig. 2.7. We simplified a single randomly generated formula ($N = 10^5$, $\alpha = 4.24$) at several fixed values of pseudo-temperature. The solid line shows for comparison the WalkSat results after a standard SP decimation (*i.e.*, $y \to \infty$): the ground state, $E = 0$, is reached as expected, after a rather small number of spin flips. The same happens after SP-Y simplifications performed at a large enough pseudo-temperature ($y > 4$); one should remind indeed that in the SAT region the optimal value for y would be infinite, and that in that limit the SP-Y recursions reduce to the SP equations. After simplification with smaller y's, the WalkSat cooling curves reach again a saturation level, which is nevertheless below the Gardner energy, unless y is too small: the threshold states of the original formula have not been able to trap the local search, even if the ground state becomes inaccessible. As we have indeed already discussed, working at finite temperature increases the probability of violating a clause when doing a spin fixing, and this is particularly evident in the SAT region where every assignment that does not satisfy some constraint should be filtered out.

The procedure is intrinsically error prone, and it will allow in general to reach only "good states", but not the true optimal solutions (the smaller the parameter y, the higher the saturation level will be). As we shall discuss in the next section, the use of backtracking partially cures the accumulation of errors at finite y: the saturation level can in fact be significantly lowered by keeping the same pseudo-temperature and introducing a small fraction of backtrack moves during the simplification. In Fig. 2.7 the data for $y = 1.5$ shows the importance of backtracking. While the run of SP-Y without backtracking has led to a plateau above Gardner energy, with the introduction of backtrack moves we find energies well below the threshold.

2.4.2 UNSAT Region

When entering the UNSAT region, the task of looking for the optimal state becomes harder. The expected presence of violated constraints in the optimal assignments really forces us to run the simplification at a finite pseudo-temperature. Unfortunately, after many spin fixings, the recursions (2.14), (2.15) stop to converge for some finite value of y before the maximum of the free energy is reached, most likely because the sub-problem has entered a full RSB phase. At this point one should switch to a 2-RSB version of SP which we did not realize, yet. Alternatively, one could try to run directly the final heuristic search (hoping that the full RSB sub-system is not exponentially hard to optimize) or more simply one may continue the decimation process by selecting the largest y for which the computation converge. We decided to implement the latter choice until either convergence is lost independently from the value of y or a paramagnetic state is reached.

In our experiments we studied several 3-SAT sample problems belonging

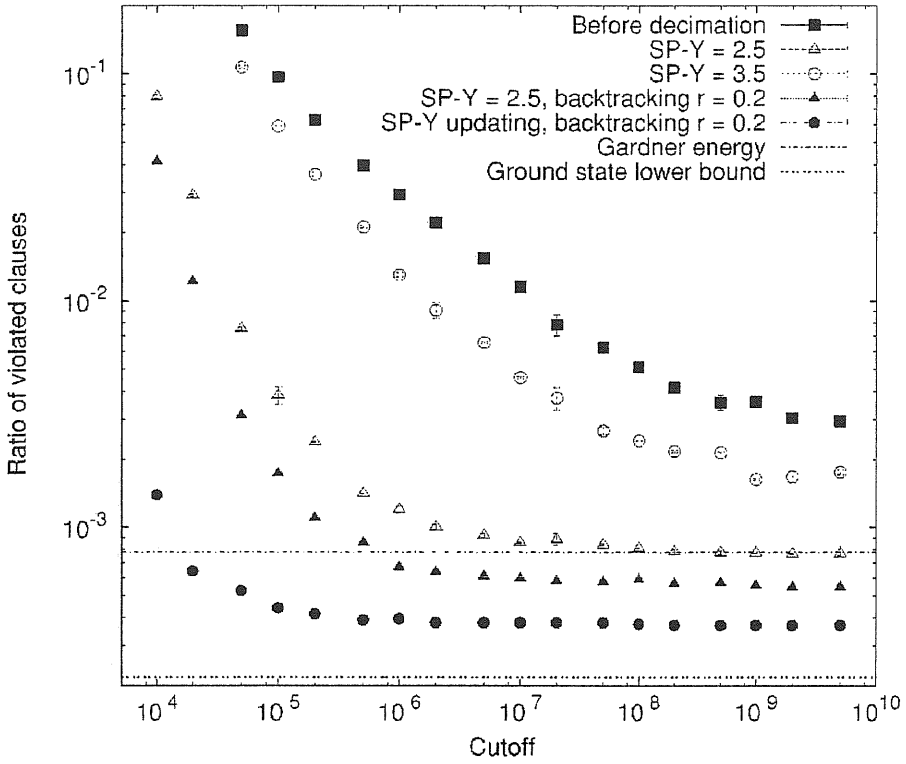


Figure 2.8: SP-Y performance in the UNSAT region (single sample with $N = 10^5$ variables and $\alpha = 4.29$). Several simplification strategies are compared; the need for backtracking is readily visible, and its introduction allows to reach energies closer to the ground state than to the Gardner lower bound.

to the 1-RSB stable UNSAT phase. We employed WalkSat as an example of standard well-performing heuristics. Although WalkSat is not optimized for unsatisfiable problems, in the 1-RSB stable UNSAT region it performs still much better than any basic implementation of SA. We observed anyway that, even after 10^{10} spin flips, the WalkSat best assignments were still quite distant from the Gardner energy, for various samples of different size and α. In Fig. 2.8 we show the results relative to many different SP-Y simplifications with various values of y and r for a single sample with $N = 10^5$ and $\alpha = 4.29$. The simplification produced always an improvement in the WalkSat performance, but, in absence of backtracking, we were unable to go below the Gardner lower bound (although we touched it in some cases: in Fig. 2.8 we show the data for a simplification at fixed $y = 2.5$; a simplification with runtime optimization of y reached the same level).

The relative inefficiency of these first attempts of simplification was not due to the threshold effect alone, but also to an extreme sensitivity to the choice of y, as pointed out by a second set of experiments making use of back-

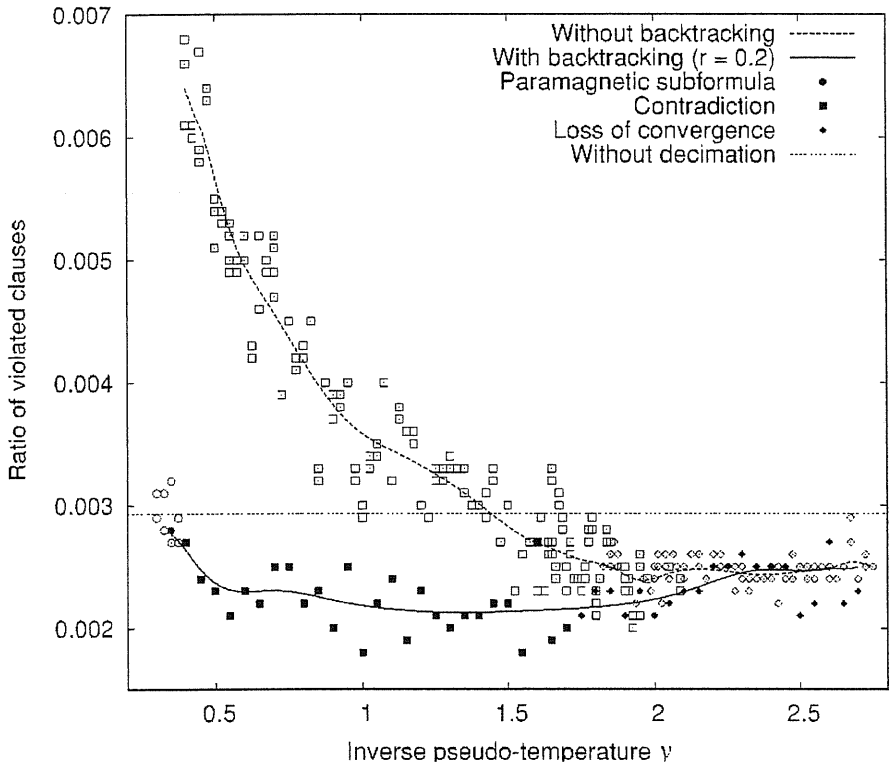


Figure 2.9: Backtracking efficiency. Many SP-Y simplifications of a single sample with $N = 10^4$ variables and $\alpha = 4.35$ have been performed at fixed but different values of pseudo-temperature; the introduction of a small fraction of backtracking moves eliminates essentially the need for a time consuming optimization of the parameter y. The empty points refer to simplifications without backtracking, the full points to simplifications with a backtracking ratio $r = 0.2$. A diamond indicates that the simplification process stopped because of loss of convergence, a circle because of finding a completely unbiased paramagnetic state, and the squares indicates that the loss of convergence happened at an advanced stage where some clause-violating assignments have already been introduced by SP-Y.

tracking. We performed first an extensive analysis of the simultaneous optimization of y and r, using smaller samples in order to produce more experimental points. After some trials, the fraction $r = 0.2$ appeared to be the optimal one, at least for our implementation, and in the small region under investigation of the K-SAT phase diagram. The data in Fig. 2.9 refers to a formula with $N = 10^4$ variables and $\alpha = 4.35$. The dashed horizontal line shows the WalkSat best energy obtained on the original formula after 10^9 spin flips. The WalkSat performance was seriously degraded when simplifying at too small values of y, but the introduction of backtracking cured the problem, identifying and repairing most of the wrong assignments. The WalkSat efficiency became actually almost independent from the choice of pseudo-temperature, whereas in absence of error correction a time consuming parameter tuning was required for optimization.

Coming back to the analysis of the sample of Fig. 2.8, the backtracking simplifications allowed us to access states definitely below the Gardner lower bound. The combination of runtime y-optimization and of error correction was even more effective: after a rather small number of spin flips, WalkSat reached a saturation level strikingly closer to the ground state lower bound than to the Gardner energy. A further valuable effect of introduction of the backtracking was the increased efficiency of the formula simplification itself: in the backtracking experiments, SP-Y was able to determine a truth value for more than 80% of the variables before losing convergence, while without backtracking, the algorithm stopped on average after only 40% of fixings.

All the samples analyzed in the previous sections were taken from the 1-RSB stable region of the 3-SAT problem, where the equations (2.4), (2.5) are considered to be exact. For $\alpha > 4.39$, the phase becomes full RSB and SP loses convergence before the free energy $\Phi(y)$ reaches its maximum from the very first step of the decimation procedure. While a full RSB version of SP would most likely provide very good results, SP-Y still can be used in a sub-optimal way by selecting the largest value of y for which convergence is reached. Numerical experiment show that indeed the performance of SP-Y are in good agreement with the analytical expectations. However, it should be noticed that in this region the use of SP is not necessary. Although the performance of WalkSat and SA can be improved by the SP simplification, the SA alone is already able of finding close-to-optimum assignments efficiently (as expected for a full RSB scenario) and behaves definitely better than WalkSat.

2.5 Conclusions

In this chapter, we have displayed the performance of SP as an optimization device and shown that configurations well below the threshold states can be found efficiently. Similar results are expected to hold also for random satisfiable instances very close to the critical point for which the combined use

of finite pseudo-temperature and backtracking could give access to the SAT optima.

It would be of some interest to analyze further improvements of the decimation strategies as well as to consider more structured factor graphs within a variational framework, in which some correlations can be put under control.

A possible application of SP-Y–like algorithms can be found in information theory: lossy data compression based on Low Density Parity Check schemes leads to optimization problems which are indeed very similar to the one discussed in this chapter.

Chapter 3

Protein as a Flexible Structure

In this chapter we introduce a two–variable coarse grained model in which we consider the network of contact interactions in the three–dimensional structure of a protein and describe changes in the flexibility of the protein when we keep only a certain subset of contact interactions. The model is mapped to a constraint satisfaction problem and solved via the Belief Propagation iterative message–passing algorithm.

As a case study, the resulting flexibility profiles are compared with the experimental temperature factors data, burial profile predictions and Molecular Dynamics simulations for the HIV Protease. The sets of contacts that drive the protein to largest flexibility are shown to describe significantly better the experimental data, confirming the notion of the HIV Protease as a flexible albeit compactly folded structure. Emphasis is given to the methodological content of the study.

The collaborative work with Michele Leone, Cristian Micheletti and Riccardo Zecchina is about to be published.

3.1 Introduction—Protein as a Flexible Structure

Although proteins are very dense macromolecules, and their density is comparable to that of macro-molecular crystals, it is not unusual that they contain regions of tens of residues that during the catalytic action can be displaced by several Å in a few nanoseconds in a concerted way. This mechanical flexibility is often of great importance for the protein catalytic function.

In this chapter we would like to address relationship between the protein flexibility and its three–dimensional structure. We propose a minimal model defined by a coarse–grained free energy function (FEF) that consists of two terms: the first contribution to the FEF comes from the contact energy of the residues in native–like state and the other term is an entropic drive to the structures with higher flexibility. This latter term counts the total number of internal degrees of freedom (DoF). The ultimate question is whether impos-

ing of flexibility constraints leads to a better agreement between predicted properties of protein and experimental data available when compared to the prediction of other coarse–grained models.

The model in its simple form cannot provide complete understanding of the relation of evolved structure and flexibility, but we hope that the results provide worthy insights in the problem. There are several ways how to generalize this approach and make it more detailed. Still, we will remain on a level as simple as possible, in order to elucidate the most general properties of the system in a transparent way.

3.1.1 The Rigidity Theory and the Model

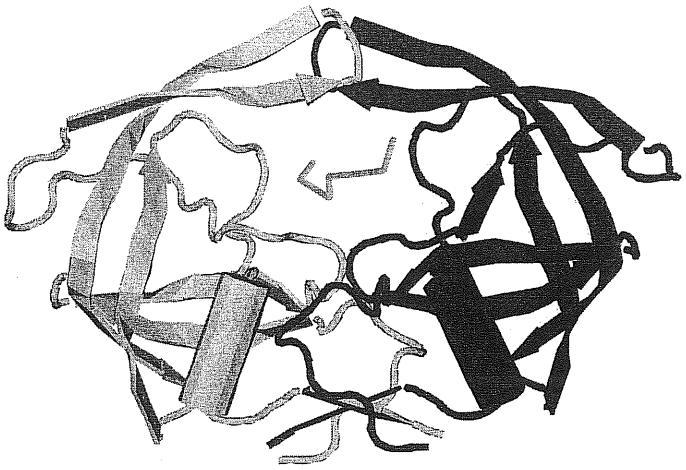
Current results of Thorpe's [59, 64, 65] group have shown that, in their model of proteins denaturation, the flexibility of protein structures increases during thermal unfolding. In the full atomic model they have located rigid and flexible regions of the protein and have followed their evolution during denaturation. We would like to continue in the direction that they have proposed but, rather then using a full atomic model, to define and solve a proper coarse–grained one, to focus only on the DoF relevant for large scale flexibility.

To represent the protein, we start from its 3D structure and at the level of single residues. Each residue is represented by its C_α atom. Only after coarse–graining we create two graphs that reflect the actual spatial structure of the protein. The first graph is the succession of backbone residues: we call it b [1]. The second graph, the contact map g, is created from the 3D structure in the customary way: two residues are in contact if the geometrical distance of their C_α atoms is smaller than 7.5Å. In this chapter we present a case study on the HIV-protease protein. Its contact map and a graphical representation is shown in the Figure 3.1.

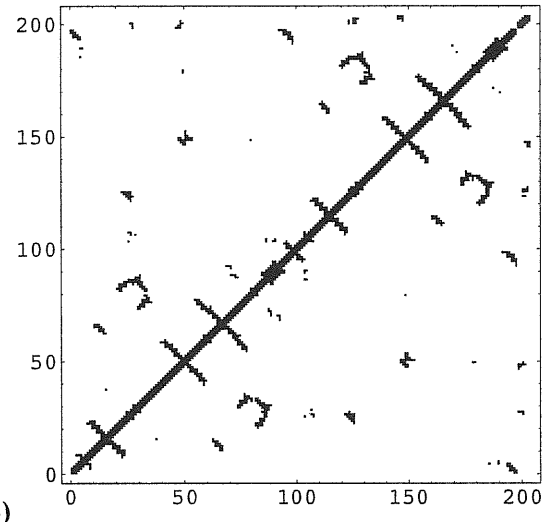
Similarly with the models of protein folding dynamics [67, 68, 69, 70], its mechanism [71], and with similarity to the models of helix–coil transition [72, 73] we consider that the residues may be in two different states: either in the *active (native) state*, that means interacting with all their neighbours in the contact map, or in the *inactive (disordered) state* when they do not interact with the neighbours on the contact map. On the other hand, the residues will interact always along the protein backbone, since we will work at moderate temperatures at which the covalent bonds cannot be broken.

For the calculation of the entropic part of the FEF we consider only the *active contacts* in the contact map, that means only the contacts with both residues in the active state, and the contacts of the backbone. This defines a subgraph of the graph g that contains only the active and hence interacting

[1] We prefer to work also with the protein backbone as with a graph, although one could imagine simpler way how to describe it. In our case it will be useful to maintain the graph representation, so that we can use similar mathematical techniques for the backbone b and the contact map g.



a)



b)

Figure 3.1: **a)** The graphical model of HIV-1 protease homodimer with an inhibitor, the secondary structure of the protein is highlighted. The central cavity in-between of its two monomers is the catalytic region of the protease, we show the placement of the inhibitor, too. The flaps above the cavity are highly mobile. **b)** The corresponding contact matrix at 7.5Å. Residues are labelled according to their position in the protein sequence. Residues $0 - 98$ form the first monomer of the protein, residues $99 - 197$ the second monomer, and finally residues $198 - 203$ belong to the inhibitor.

residues. We call the subgraph g^{int}. In order to evaluate the entropic part of the FEF, we have to calculate the flexibility of the structure represented by the graph g^{int} and the backbone in an efficient way.

It is the Pebble Algorithm of Thorpe and co-workers [65] that allows us to sample the space of all possible subsets of active/inactive residues of a given protein in fast and schematic way.

In the following Section, we will present the model Hamiltonian that we will use to extract those subsets of active residues that lead to the largest flexibility, with the aim of comparing the subsets to structural data.

3.1.2 Free Energy Function of the Flexible Protein

If we denote the active/inactive state of the residue by a spin–like variable $s_i = 1(0)$ for active (inactive) residue, we can write the energetic part of the FEF as

$$H^{nat} = -\sum_{(i,j)\in g} s_i s_j. \tag{3.1}$$

Where the superscript stands for *native*. The topology–based energetic term resembles the Hamiltonian of the Gō model which has been proposed for description of unfolded/native transition in proteins [74]. The term counts number of active links in the contact map.

The entropic part of the FEF takes form $-\frac{\nu}{\beta}H^{DoF}$ where H^{DoF} is number of floppy modes and ν is an appropriately chosen adimensional constant. To evaluate the entropy term we have to calculate the flexibility of the structure with the activities of residues assigned by the set of active residues $\{s_i\}$. The FEF takes form

$$H = H^{nat} - \frac{\nu}{\beta}H^{DoF}, \tag{3.2}$$

where $\nu > 0$ is used to balance the energetic and the entropic term. We have chosen the value of ν from experimental data [75, 76] to be $\nu \simeq 0.5$.

Finally, as we will be interested in finding various assignments of active residues $\{s_i\}$ within the $3D$ structure, we add to the statistical weight external field term which fixes the residue variables s_i to some desired target sequence $\{\bar{s}_i\}$. This will allow us to select the residue contacts that are actually present in the structure:

$$H^{overlap} = 4(\frac{1}{2} - \bar{s}_i)(\frac{1}{2} - s_i) = 2\delta(s_i;\bar{s}_i) - 1. \tag{3.3}$$

The total statistical weight is then:

$$H = H^{nat} - \frac{\nu}{\beta}H^{DoF} - \frac{\lambda}{\beta}H^{overlap}, \tag{3.4}$$

where the parameter λ will be chosen large enough in order for residue site variables to be chosen equal to the desired target sequence ones.

3.1.3 The Pebble Algorithm for the Calculation of Flexibility

Imagine a 3D structure made of joints and sticks. Our question is if the structure is rigid, and if not, how much it is flexible, how many internal DoF it has.

The Pebble Algorithm [65] counts the degrees of freedom in a very direct and intuitive way. Each joint has in 3D space three DoF and the number of the DoF can be decreased only by some mechanical constraints. The constraint in our case is the existence of a stick that connects the joint to some other joint and hence it fixes distance between them (We will discuss later on why each stick removes only one DoF).

Such a stick connects two joints so we have to decide somehow, to which joint we decrease the number of DoF. Our decision can be expressed by placing an arrow to the stick and making it directed. We can orient all the sticks in the structure. Given such an orientation, the number of degrees of freedom of each joint i equals $\max\{0, 3 - n_i\}$, where n_i is the number of sticks that point to it. The maximum ensures that the number of DoF is not negative—if there are three or more constraints on the joint, it has no DoF left. Still the orientation of the sticks cannot be done in a random way, we have to choose the correct orientation among all possible. The criterion of correctness is minimization of the overall number of floppy modes. We may write

$$H^{DoF} = \min_{\{n_i\}} \sum_i \max\{0; 3 - n_i\} - 6, \tag{3.5}$$

where we subtract the six rototranslational degrees of freedom of the motion of the protein as a whole.

Going back to the model of protein, we interpret the joints as residues (C_α atoms) and the sticks as their contacts. Each C_α atom and consequently the residue is represented by a single point with at most three DoF. The contacts of the protein backbone represent the peptide bonds. In the peptide bond there are two angles partially free to rotate and hence we consider the backbone contact to fix only the distance of the residues and to remove one DoF only. Similarly for the links of the contact map, we consider that the interaction fixes only one degree of freedom.

To count properly the number of links pointing to each residue we will place on each link (i, j) an auxiliary link variable x_{ij} with two states that correspond to the direction of the link. We define the value of the link variable $x_{ij} = 1$ ($x_{ij} = 0$) if the link is directed towards the residue j (i). Clearly $x_{ij} = 1 - x_{ji}$. The formula (3.5) may be rewritten in the means of link variables as

$$H^{DoF}(\{\widetilde{x_{ij}}\}) = \min_{\{x_{ji}\}} \left(\sum_i \max\{0; 3 - \sum_{j \in g^{int}(i) \cup b(i)} x_{ji}\} - 6 \right), \tag{3.6}$$

where we call $\{\widetilde{x_{ji}}\} = \{\widetilde{x_{ji}}(\{s_i\})\}$ the configuration of link variables that minimizes the H^{DoF} and $g(i)$ $(b(i))$ is the set of all neighbours of site i in the contact map (backbone). $H^{DoF}(\{\widetilde{x_{ij}}\})$ is then the actual number of floppy modes of the protein structure with all residues assigned to be either active or inactive. The complete FEF is then

$$H = -\sum_{(i,j)\in g} s_i s_j - \frac{\nu}{\beta}\Big[\sum_i \max\{0; 3 - \sum_{j\in g(i)} s_j s_i \widetilde{x_{ji}} - \sum_{j\in b(i)} \widetilde{x_{ji}}\} - 6\Big] - \\ - \frac{\lambda}{\beta}\sum_i \Big[(\frac{1}{2} - \bar{s_i})(\frac{1}{2} - s_i)\Big], \tag{3.7}$$

where we have expressed the sum over the graph g^{int} by the sum over full graph g but using the residue variables s_i.

3.1.4 Separation of the Calculation in Two Parts

As previously stated, we want to analyze the subsets of active residues that lead to highly flexible structures. In order to do so, the procedure is naturally splitted in two steps. To evaluate the free energy, one has to minimize equation (3.5) first for a given assignment $\{s_i\}$ in order to find the actual flexibility of the structure. While sampling the set of active residues (fixed in number) providing the maximal flexibility one has to maximize similar term in (3.7) over all possible assignments $\{s_i\}$.

Hence we can separate the problem in two parts. We will study the configuration space with the canonical weight given by the energy:

$$H = H^{nat}(\{s_i\}) - \frac{\nu}{\beta} H^{DoF}(\{\widetilde{x_{ij}}\}, \{s_i\}), \tag{3.8}$$

where in order to find the configuration $\{\widetilde{x_{ij}}\}$ we will sample an auxiliary Hamiltonian in the form:

$$H^{flexibility} = \frac{\mu}{\beta}\Big[\sum_i \max\{0; 3 - \sum_{j\in g(i)} s_j s_i x_{ji} - \sum_{j\in b(i)} x_{ji}\} - 6\Big] - \\ - \frac{\lambda}{\beta}\sum_i \Big[(\frac{1}{2} - \bar{s_i})(\frac{1}{2} - s_i)\Big], \tag{3.9}$$

with $0 < \mu \ll \lambda$ and μ large enough to minimize the flexibility term and to find the configuration $\{\widetilde{x_{ij}}\}$.

3.1.5 Random Sampling

Although the complete solution to the two coupled problems is possible, both using a double Belief Propagation algorithm and/or combined method where

Monte Carlo is used for the real FEF and the Belief Propagation is used for the auxiliary Hamiltonian, we are going to sample the configuration space in a simpler and clearer manner. As we are interested mainly in the influence of the flexibility on the statistics of subsets of active residues, we work at a fixed total Gō contact energy (number of contacts).

We generate a large enough random ensemble of subsets with equal contact energy and on this ensemble we study the influence of the flexibility on the statistics of the subsets of active residues. Namely we are going to compare prediction on the number of DoF at an arbitrary residue with experimental temperature factors (see the formula (3.37) for a definition of temperature factors).

The number of degrees of freedom is the quantity that in our model is closest to the amplitude of atomic motions which is estimated by the temperature factors, although we see only the number of DoF left on the residue and hence the dimensionality of the motion.

3.1.6 The Belief Propagation for the Calculation of Flexibility

As our ensemble contains the subsets of active residues that lead to the same contact energy, we have only to evaluate the flexibility term. The auxiliary flexibility Hamiltonian may be expressed as a sum

$$H^{flexibility} = \sum_{i=1}^{N} h_i, \tag{3.10}$$

where h_i is defined as

$$\begin{aligned} h_i = \; & \frac{\mu}{\beta}\Big[\max\{0; 3 - \sum_{j\in g(i)} s_j s_i x_{ji} - \sum_{j\in b(i)} x_{ji}\} - \frac{6}{N}\Big] - \\ & - \frac{\lambda}{\beta}\Big[(\frac{1}{2} - \bar{s_i})(\frac{1}{2} - s_i)\Big]. \end{aligned} \tag{3.11}$$

Since the Hamiltonian contains only nearest neighbours interactions we may write a Belief Propagation procedure that solves it within Bethe approximation [57] following the cavity approach of Barré *et al.* [66].

The *Belief Propagation* is a self–consistent local message passing algorithm originally devised in the field of pattern recognition, information and coding theory, and artificial intelligence.

In statistical physics it corresponds to Bethe approximation, *i.e.*, to the assumption that the graph on which BP is performed (Figure 3.2) is a tree. In the last years it has been widely shown [61, 62] that BP works very well also in graphs that are only locally tree–like, and even in the presence of more regular or dense geometrical structure in cases where correlation decay fast enough with distance. To this regard, we should say that the BP approach

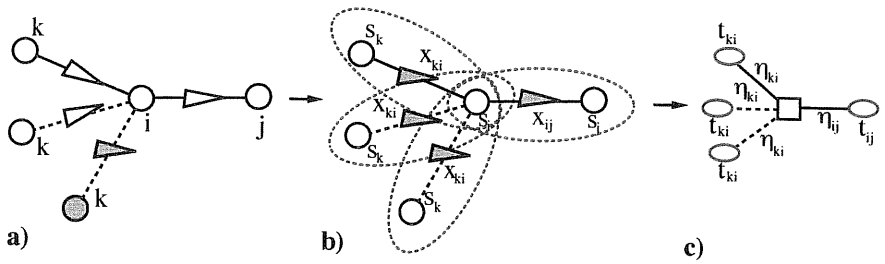


Figure 3.2: **a)** The Pebble game representation. The residues i are represented by balls, active residues are coloured grey. There are two kinds of links: the backbone links (thick lines) and the contact map links (dashed lines). The orientations of links are shown as arrows. Orientation of links between inactive residues is arbitrary (grey). **b, c)** Dual-like transformation of the contact graph: new aggregated Potts variables are placed on original links. The resulting structure is a factor graph where function nodes are displayed as squares. Nodes retain the FEF constraint plus a hard core compatibility constraint stating that variable s_j is common to several t's and must therefore take the same value within them. This leads to a selection of allowed values for macro-variables that insist on the same function node. Compatibility values can easily be found from inspection of Table 3.1. **c)** Cavity/BP messages flow from function nodes to macro-variables. By construction, function nodes are attached to original residues sites and macro-variables to links, so that their degree is fixed to 2. The energy term in equation (3.8) has to be split evenly among all function nodes of the "dual" graph.

t_{ij}	0	1	2	3	4	5	6	7
s_i	0	0	0	0	1	1	1	1
s_j	0	0	1	1	0	0	1	1
x_{ij}	0	1	0	1	0	1	0	1

Table 3.1: The coding of residue and link variables to aggregated variables.

can be generalized through a standard Cluster Variation Method (CVM) procedure. It might be the case that this approach will be useful in the case of protein graphs. However, as stated in the chapter introduction, this line of work goes beyond the aim of the present analysis.

In order to write down the explicit message passing procedure it is useful to merge the residue variables $\{s_i, s_j\}$ and link variables $\{x_{ij}\}$ into aggregated Potts like macro variables $\{t_{ij}\}$. We will do it in the following way, shown and explained more in details in Figure 3.2bc: We take a link between residues i, j. It connects spins s_i and s_j, its spin variable is x_{ij}. These three spins will be considered as one aggregate variable t_{ij} with 8 possible states (see Table 3.1).

Starting from the flexibility Hamiltonian (3.10) one would like to efficiently compute the partition function

$$Z_\lambda^{flexibility} = \sum_{\{t_{ij}\}}{}' \exp\left\{ -\beta H^{flexibility} \right\} \tag{3.12}$$

and the equilibrium configurations Boltzmann probabilities as

$$P(\{t_{ij}\}) = \frac{1}{Z_\lambda^{flexibility}} \exp\left\{ -\beta H^{flexibility} \right\} \tag{3.13}$$

where the sum in the partition function runs over all compatible configurations of $\{t_{ij}\}$.

From equation (3.13), probability distributions $\{P_{ij}(t_{ij})\}$—the link probability, and $\{P_i(\{t_{ij}\})\}$—the joint probability distribution of all collective variables sharing the site variable i in the original graph—, may be expressed. These quantities are defined as marginalizations of (3.13) over all other variables.

Bethe Ansatz amounts to a variational hypothesis where the protein contact graph $b \cup g$ is treated as if it was a bipartite tree. In this case it is easy to prove (*e.g.*, by induction) that the following factorization property holds:

$$P(\{t_{ij}\}) = \frac{\prod_i P_i}{\prod_{ij} P_{ij}^{c-1}} \tag{3.14}$$

where c is the degree of the collective variables and is always equal to 2 in our case, and local probabilities are trial functionals whose best form has to be found variationally as the one that minimizes the free energy potential.

Inserting Ansatz (3.14) in (3.13) and imposing self–consistent marginalization rules for local probabilities and minimization of the Gibbs free energy potential, it is possible to write a set of closed equations for messages $\{\eta_{ij}(t_{ij})\}$ describing a local probability flow on each link of the bipartite graph of Figure 3.2. Messages are just a functional way to split up probability contributions over all the network links, such that one can further write, at convergence of the self–consistent equations (η^*), beliefs as:

$$P_{ij}(t_{ij}) \propto \eta^*_{ij}(t_{ij})\eta^*_{ji}(t_{ij}) \tag{3.15}$$

and

$$P_i(\{t_{ji}\}) \propto \prod_{j\in g(i)\cup b(i)} \eta^*_{ji}(t_{ji}) \exp\{-\beta h_i\}. \tag{3.16}$$

Within the region of validity of Bethe approximation and in the region of the model phase space where no clustering phenomenon on the organization of thermodynamic states is present [2], probabilities for each macro-variable to take any allowed value can therefore be written as the fixed point solutions of an iterative procedure that solves a system of non-linear equations. Derivation of these equations can be done via an iterative computation of the partition function of the model, as stated above, but here we will state only the results. In order to do so, it is useful to separate the terms of $\exp(-\beta h_i)$ that contain only the variables local to the residue i and the rest where also another quantities may occur,

$$\begin{aligned} \exp(-\beta h_i) = \; & \exp\left\{\lambda(\frac{1}{2}-\bar{s_i})(\frac{1}{2}-s_i)+\frac{6\mu}{N}\right\}\times \\ & \exp -\mu \max\left\{0; 3-\sum_{j\in g(i)} s_j s_i x_{ji} - \sum_{j\in b(i)} x_{ji}\right\}, \end{aligned} \tag{3.17}$$

$$\exp(-\beta h_i) = \; T(s_i,\bar{s_i}) \exp\{-\mu h_i^{DoF}\}. \tag{3.18}$$

The last formula defines quantities $T(s_i,\bar{s_i})$ and h_i^{DoF}. We will call $T(s_i,\bar{s_i})$ the term that depends only on the state of the variable s_i. For the belief $\eta_{ij}(t_{ij})$ that on the link (i,j) the state variable takes value t_{ij} in absence of the second constraint connected to t_{ij} through site j, the system of self–consistent equations (one for each link of the bipartite graph) reads as

$$\eta_{ij}(t_{ij}) = \frac{1}{C_{ij}} T(s_i,\bar{s_i}) \sum_{\{t_{ki}\}}{}' \prod_{k\in g(i)\cup b(i)\setminus j} \eta_{ki}(t_{ki}) \exp\{-\mu h_i^{DoF}\}. \tag{3.19}$$

Again, we write the prime over the sum because not all configurations of $\{t_{ki}\}$ are allowed. Only those where the quantity s_i is fixed in all t_{ki}'s to

[2] The interested reader is referred to [26, 24, 60, 58] for discussion and extension of the BP algorithm to such cases, which however are not expected to apply in the present model.

the same value. System (3.19) can be implemented algorithmically in what is known as the message–passing Belief Propagation procedure.

Let us explain all the terms in the (3.19) with the help of Figure 3.2c. The out-coming belief η_{ij}, is expressed in terms of all incoming believes η_{ki}. Their 'convolution' is then weighted by the on-site weight of the given configuration $\exp\{-\mu h_i^{DoF}(t_{ki},\ k \neq j)\}$ and by a pre-factor $T(s_i, \bar{s_i})$. The belief η_{ij} is then normalized by C_{ij},

$$C_{ij} = \sum_{t=0}^{7} \eta_{ij}(t). \tag{3.20}$$

We have devised a fast message passing procedure that leads to self-consistently assigned values of $\eta^*_{ij}(t_{ij})$ (fixed point under the iteration of (3.19)) for all i, j. See the Appendix for more details on the scheme.

Once having self-consistent η^*'s we can calculate all thermodynamic properties of the system. Among them we will be most interested in the flexibility. As we will keep μ very large, the mean value of the flexibility term of the auxiliary Hamiltonian will give the actual flexibility for a given sequence $\{s_i\}$.

First we have to calculate probability of a given site i to be in the active state. To do so, we have to introduce probability of finding configuration of $\{t_{ki}\}$ around the site i. Similarly to the expression for the belief η_{ij} this probability is (after convergence):

$$P_i(\{t_{ki}\}) = \frac{1}{C_i} \prod_{k \in g(i) \cup b(i)} \eta^*_{ki}(t_{ki}) \exp\{-\beta h_i\}. \tag{3.21}$$

where C_i is the normalization constant that can be expressed as

$$C_i = \sum_{\{t_{ki}\}} P_i(\{t_{ki}\}). \tag{3.22}$$

Once knowing the probability of the state of a node $P_i(\{t_{ki}\})$, we can calculate the mean flexibility as the ensemble average. If we take the average number of degrees of freedom on site i, $\langle h_i^{DoF} \rangle$, we may write the mean total flexibility as:

$$H^{DoF}(\{\widetilde{x_{ij}}\}, \{s_i\}) = \frac{1}{N}(\sum_i \langle h_i^{DoF} \rangle - 6). \tag{3.23}$$

The number of degrees of freedom on the residue i is

$$\langle h_i^{DoF} \rangle = \sum_{\{t_{ji}\}} \prime P_i(\{t_{ki}\}) \max\{0; 3 - \sum_{k \in g(i)} s_k s_i x_{ki} - \sum_{k \in b(i)} x_{ki}\}. \tag{3.24}$$

In the case when one is interested in knowing the probability distribution of a residue site variable, the result can be computed simply via marginalization of $P_i(\{t_{ij}\})$'s. Indeed, since there are many $P_i(\{t_{ij}\})$ that contain the same site information, marginalization can be used as an internal consistency check on

BP results. We show in the Appendix how one can set up a fast algorithmic way to evaluate the update rule (3.19) in order to find the self–consistent values η^*. A similar scheme is used for the evaluation of (3.24).

3.2 Results for the HIV-Protease

As a concrete test for our analysis we have chosen the HIV-1 Protease, a very well studied protein with many available experimental and theoretical results. Its steric structure is known to a high detail and the experimental values for the temperature factors are available [63]. It also contains mobile regions, since during the catalytic reaction some regions change non trivially their conformation. On the other hand, other regions are relatively rigid. A graphical model of the protease is shown in Figure 3.1. The protein is a homodimer, in between of the two monomers there is the cavity with the active region. Each monomer has the length of 99 residues. In our calculation we have worked with the protein in complex with a simple peptide–like inhibitor of 6 amino–acid residues (shown in the cavity).

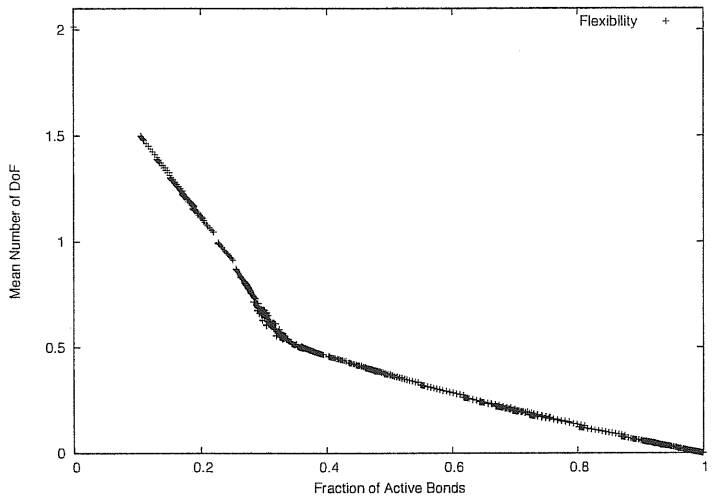
Together with experimental data, there is also data obtained from Molecular Dynamics analysis [77] and well established coarse–grained model for protein dynamics—Gaussian Model [80, 78, 79].

The 3D structure of the protease defines in our idealization the two graphs g and b. The possible subsets of active residues are generated as assignments of residues to be active or inactive (with close similarity to the HP model). For any generated sequence we can calculate its flexibility (3.23) and the profile of DoF. Our goal will be to show that for the subsets that lead to high flexibility we find very good correlation with temperature factors.

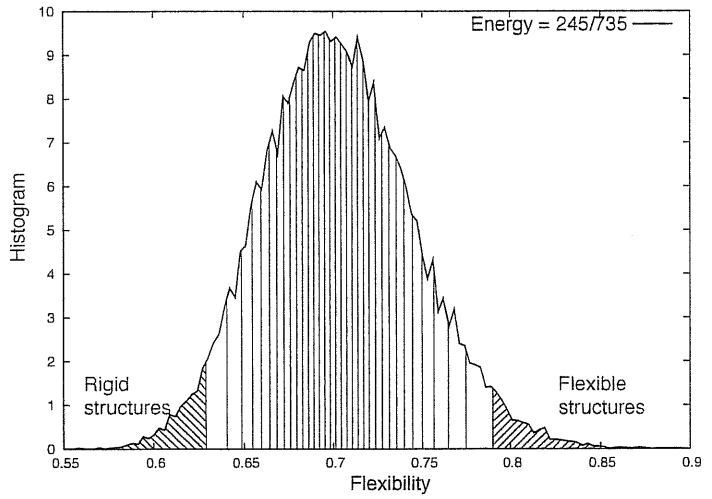
In our calculation we strictly followed the scheme of the previous section. First, we generated approximately 28000 [3] random subsets of active residues with an appropriately chosen energy. We have chosen three energies–one which is specified by a pronounced kink in the flexibility vs. energy curve and other two energies that correspond to much less (more) connected network, respectively. The dependence of flexibility on contact energy is shown in the Figure 3.3a. The selected energy levels correspond to the 165, 245, 485 active bonds present in the network out of 735 links in total.

Following the method of the previous section, we sampled the flexibility of all subsets of active residues at a given energy. Then we have created a histogram of flexibility values in the ensemble and we have chosen the subsets that lead to the most flexible and the most rigid structures for comparison

[3] We have done the analysis also on smaller ensembles of 2000 and 16000 different subsets. The results did not change when increasing the size of the ensemble, so we consider the ensemble representative. As the algorithm is quite efficient we have chosen as large ensemble as possible. Bigger number of sequences would cause problems with the large RAM needed for calculation of two–point correlation matrix.



a)



b)

Figure 3.3: **a)** The dependence of average flexibility of the ensemble of subsets of active residues with a given energy. Energy is equivalent to the fraction of active links. **b)** In the histogram of flexibilities the stripes that are used for the analysis of similarity to temperature factors are shown. Each of the stripes corresponds to 800 different subsets, all together there are 28000 subsets in the ensemble (19200 for energy 165/735). We show data for energy 245/735, other data are similar yet centred around different mean value of flexibility. Flexibility decreases with the growing number of bonds.

with the temperature factors. See Figure 3.3b. From the histogram of flexibility we have taken out three ensembles of subsets. At any given energy we have created ensemble containing the subsets that induce the most rigid and the most flexible structures. As a control ensemble we have taken the complete set of subsets.

For each subset from either ensemble we have counted the number of DoF of a given residue, using the Belief Propagation scheme shown in the Appendix. We call the *DoF profile* a set of the values for all residues. As the DoF profile gives the mean dimension of the residue fluctuations we can compare it with the experimental temperature factors that are also a measure of residue fluctuations.

The comparison was done in the following way. For each subset of active residues we calculated correlation between the DoF profile and the temperature factors. Then we plotted the distribution of these correlation coefficients, see Figure 3.4.

We immediately see that the subsets that drive the more flexible structures correlate to temperature factors much better and hence they describe better the actual structure of the protein. The average DoF profile for the three ensembles is shown in Figure 3.5, together with the corresponding temperature factor profile.

What is the level of prediction? How well does this method do when compared to other currently available methods? To answer these questions we have compared our results with temperature factors prediction extracted from information on the burial profile of the protein. The *burial profile* is a sequence of numbers that give the connectivity of each residue, the number of its neighbours within some cut-off distance (7.5 Å). It is well know that the prediction of the temperature factors from the burial profile can be very accurate (with the linear correlation coefficient of 0.6) and may be even improved within a certain scheme [81].

The burial profile in our scheme corresponds to the contact map g, so we compare these two properties: first, we show the correlation between the predicted DoF profile and the burial profile of the protease. Second, we can compare matrices of two-point correlations with the contact matrix.

The correlation with the burial profile (see Figure 3.6) is very high for the subsets of active residues that induce the most flexible structures. That on one side means that the prediction of our model is not much richer, but on the other side it shows that the protein topology is optimized for its function. The flexibility is closely linked to the 3D structure.

To show that there is more information in the DoF profile than in the burial profile, we evaluated the two-point correlation functions and compared them to the results from Molecular Dynamics. If the two-point correlation functions show only the features of the contact matrix, we cannot expect more information to be stored in the DoF profile than in the burial profile. Indeed, correct pair-correlations may be obtained only by considering the three-dimensional

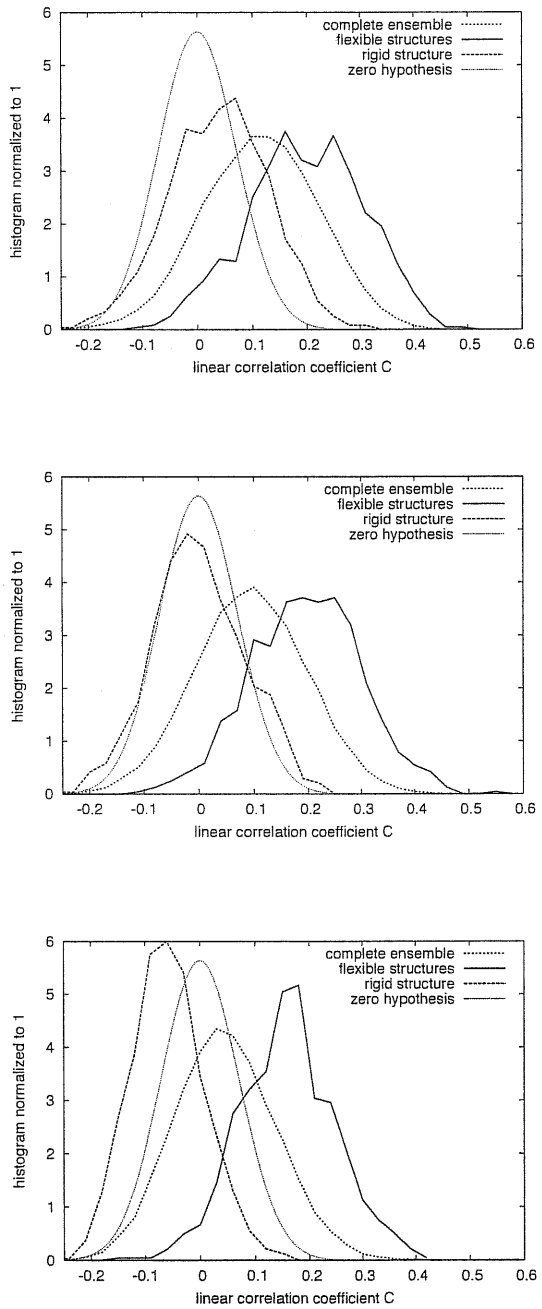


Figure 3.4: The histogram of DoF-profile–to–temperature-factors correlations. From top to bottom: the low, medium, and large energy (number of active bonds). One clearly sees much better correlation of more flexible sequences. The smooth curves show zero-hypothesis for the correlation coefficient.

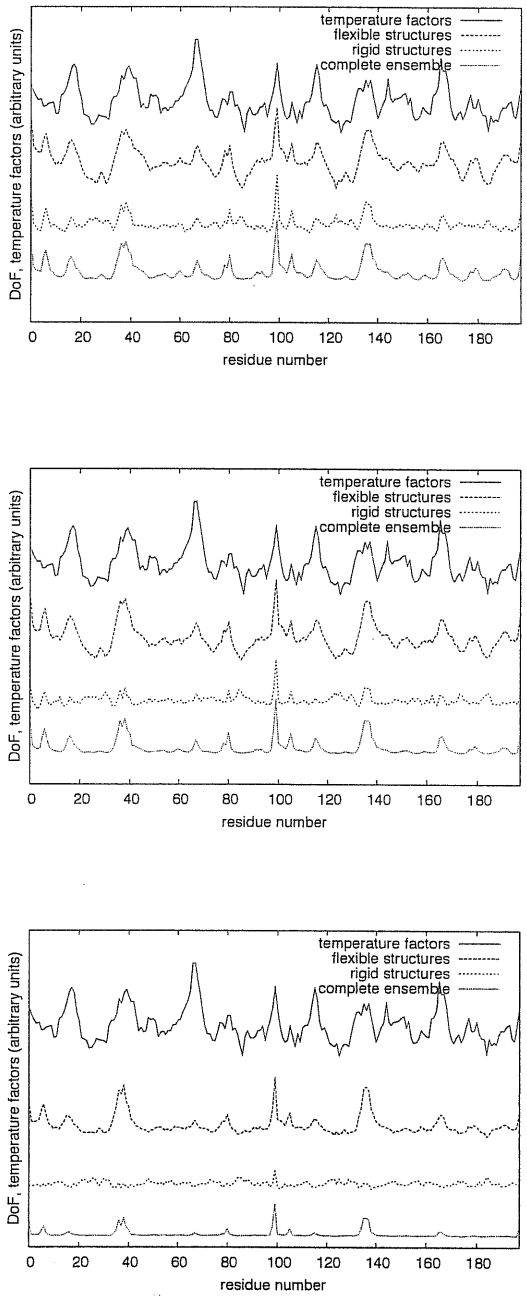


Figure 3.5: The comparison of mean DoF profile with the experimental temperature factors for the rigid, average and flexible subsets of active residues. From top to bottom: number of bonds 165, 245, 485.

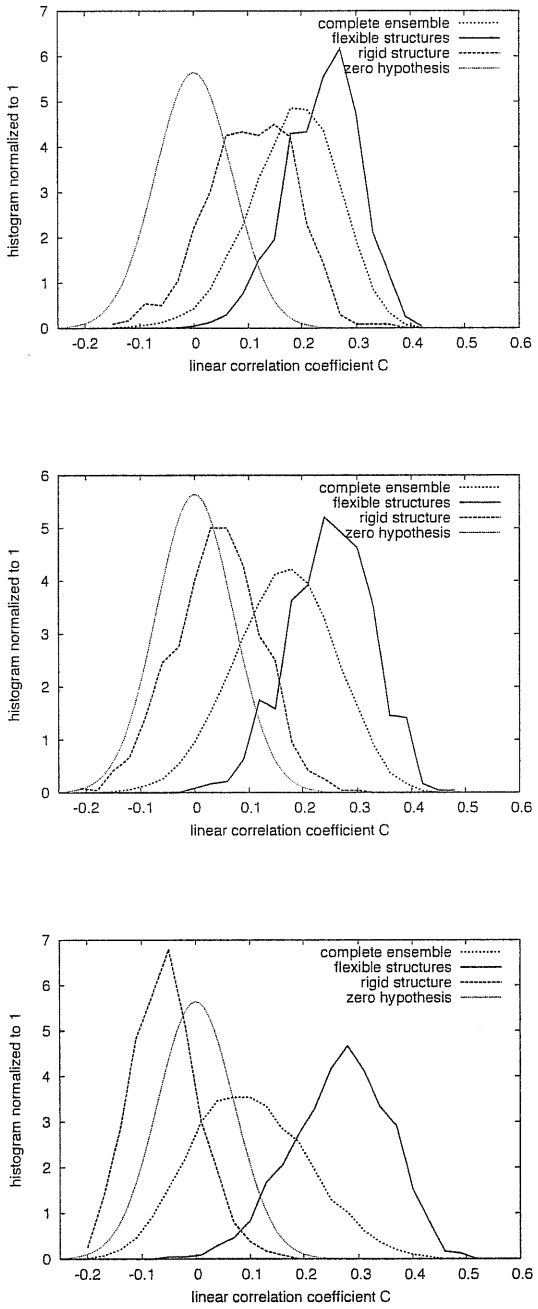


Figure 3.6: The histogram of correlations of DoF profile to the inverse of the burial profile. From the top: the low, medium, and large energy (number of active bonds). The smooth Gaussian curves show zero-hypothesis for the correlation coefficient. The most flexible subsets correlate with the burial profile much better than those rigid. The correlation increases with the increasing number of bonds.

orientation of the contacts.

We computed two-point correlations as the linear correlation between the number of DoF at sites i and j. We have checked that also other correlation measures (Kendall's τ) give equivalent results. A reason for use of the linear correlation coefficient was the partially discrete nature of number of DoF: we may expect that many subsets of active residues induce similar or equal values of h_i^{DoF}.

Another possible measure of relationship of two residues would be the pair-correlation coefficient of the fluctuations of the number of DoF of the two residues. For completeness we have considered also this quantity, but we have found no improvement to the direct pair-correlations of the number of DoF. The number of degrees of freedom is a good measure of the residue mobility.

We evaluated similarity of the two-point correlation matrices with the two-point fluctuation correlations obtained from Molecular Dynamics [4]. In order to discard information stored in the burial profile, we have neglected the correlations between the residues i and j if they are either directly in contact or in the backbone sequence ($i \in b(j) \cup g(j)$), or their neighbours in the backbone sequence ($m \in b(i)$) are in contact with some other residue ($m \in b(j) \cup g(j)$). Therefore we have subtracted trivial correlations that come only from the burial profile.

The correlations are not very high but still significant. For the smallest energy (165/735) at which there is smallest number of bonds we have the correlations of $\tau_{flexible} = 0.10$ ($\tau_{rigid} = 0.00$) for the 'flexible' ('rigid') subsets of active residues. With the increasing number of bonds the difference of correlation coefficients is less and less pronounced. For energy 245/735 we have $\tau_{flexible} = 0.07$ and $\tau_{rigid} = -0.01$, for the highest energy then $\tau_{flexible} = -0.01$ and $\tau_{rigid} = -0.01$. The last results reflects the fact that the mechanical structures are over-constrained once there is too many bonds present.

The resulting values clearly show good correlation for the subsets that lead to flexible structures and very poor correlation for the subsets that impose the most rigid structures. Still, we do not reach the correlation obtained with the Gaussian Model ($\tau = 0.43$). The data for the best correlating case (flexible sequences at the energy 165) are compared with the Molecular Dynamics data in the Figure 3.7.

3.3 Conclusions

In this work we have introduced a coarse–grained model of protein flexibility in which the corresponding Hamiltonian represents a set of steric and en-

[4] The correlation from Molecular Dynamics are correlation of fluctuations around the average position of the atom, so they are related to our two–point correlation functions indirectly.

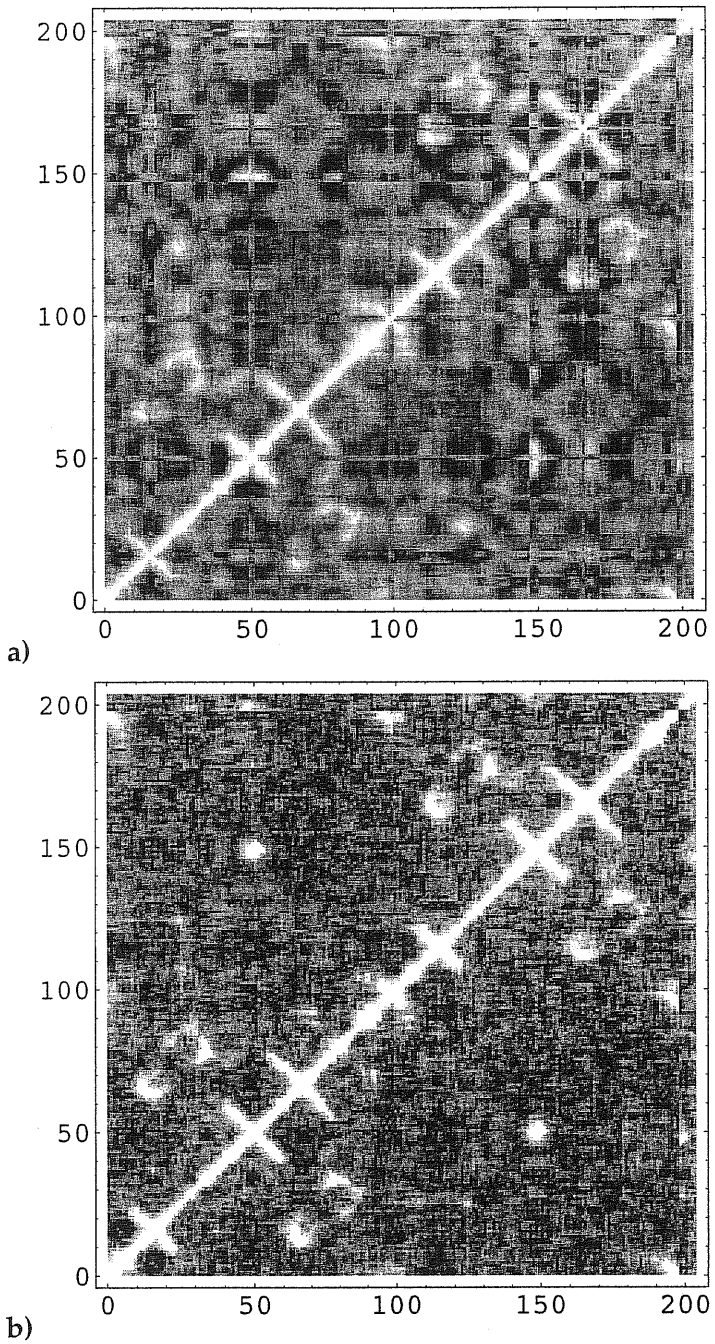


Figure 3.7: The two–point correlations from Molecular Dynamics [77] (**a**)) and the two-point correlations of the DoF profile (**b**)). The close correspondence is present for small number of bonds and for the 'flexible' subsets (energy = 165/735, flexible ensemble). On the horizontal and vertical axes are the residue numbers. We show also the residues of the inhibitor (last 6 rows/columns).

tropic constraints that can be efficiently solved via a local message–passing procedure.

The central idea underlying this approach lies in the current of research aiming to define a hierarchy of models that goes from the complete microscopic description to successive levels of coarse–graining. In this multi-scale approach, physical parameters resulted from our level are used as input to tune the next level of models in a coherent way. The final goal is that of explaining the emergence of relevant behaviour from first principles not via a direct and unfeasible brute force approach, but through a successive model–refinement procedure.

The approach is very general. Indeed, the presented work focused more on the presentation of the methodology than on direct applications. Nevertheless, a case study of practical importance (the flexibility study of the HIV–protease) was presented. In this last case, we showed that among all possible subsets of active residues, those that impose the most flexible structures, describe much better the properties of the real protein. On the other side, the 'rigid' subsets describe the structure rather poorly. Unfortunately, our tool is not powerful enough to predict the sequence of the protein. We believe that this is caused by simplicity of the model. Indeed there are many ways how to extend the model, starting from accounting of C_β atoms, calculating the contact energy with consideration of actual interacting amino-acid residues or passing to a full atomic model of the protein. All these extensions are feasible, but go beyond the intentions of the authors.

Appendix: Calculation of the Flexibility by the Belief Propagation

In this section we derive an algorithmic way to evaluate the formulae (3.19), (3.21), and (3.24).

We start with the central equation (3.19) and we show the way how to reach a self-consistent solution of the coupled equations for the complete set of $\{\eta_{ij}\}$.

In doing so, we have to separate two branches of calculation, indeed the messages along the protein backbone edges have slightly different nature than the messages sent along the contact map edges. Hence we evaluate formula (3.19) separately for the two cases.

Let us start with the case when the link (i, j) is in the backbone graph. In that case we can rewrite (3.19) in the form:

$$\eta_{ij}(t_{ij}) = \frac{T(s_i, \bar{s_i})}{C_{ij}} \sum_{\{t_{ki}\}}{}' \prod_{k \in g(i)} \eta_{ki}(t_{ki}) \times \prod_{k \in b(i) \setminus j} \eta_{ki}(t_{ki}) \exp\{-\mu \max\{0; 3 - \sum_{k \in g(i)} s_k s_i x_{ki} - \sum_{k \in b(i)} x_{ki}\}. \tag{3.25}$$

First we evaluate the formula in the case when $t_{ij} = 0$. That immediately implies that $s_i = s_j = 0$, $x_{ij} = 0$, and $x_{ji} = 1$. Hence the formula (3.25) simplifies to:

$$\eta_{ij}(0) = \frac{1}{C_{ij}} T(0,\bar{s_i}) \sum_{\{t_{ki}\}}{}' \prod_{k\in g(i)} \eta_{ki}(t_{ki}) \times \prod_{k\in b(i)\setminus j} \eta_{ki}(t_{ki}) \exp\{-\mu \max\{0; 2 - \sum_{k\in b(i)\setminus j} x_{ki}\}\}. \tag{3.26}$$

We see that depending on the value of x_{ki}, where k is the other backbone neighbour of i than j (if there is any) the maximum in the exponent may take two values, either 2 or 1. The value is 2 iff $x_{ki} = 0$ and hence iff $t_{ki} \in \{0, 4\}$. Similarly, the value is 1 iff $t_{ki} \in \{1, 5\}$. The values of all contact map messages are limited by $s_i = 0$ only, hence $t_{ki} \in \{0, 1, 4, 5\}$.

Using this information we may write:

$$\eta_{ij}(0) = \frac{T(0,\bar{s_i})}{C_{ij}} \prod_{k\in g(i)} [\eta_{ki}(0) + \eta_{ki}(1) + \eta_{ki}(4) + \eta_{ki}(5)] \times \Big(\prod_{k\in b(i)\setminus j} [\eta_{ki}(0) + \eta_{ki}(4)]e^{-2\mu} + \prod_{k\in b(i)\setminus j} [\eta_{ki}(1) + \eta_{ki}(5)]e^{-\mu} \Big). \tag{3.27}$$

It will show very useful to define the following products for messages along backbone graph b:

$$\begin{aligned}
B_{i,V}(s_i = 1) &= \prod_{k\in b(i)\setminus V} [\eta_{ki}(2) + \eta_{ki}(3) + \eta_{ki}(6) + \eta_{ki}(7)], \qquad (3.28)\\
B^{+}_{i,V}(s_i = 0) &= \prod_{k\in b(i)\setminus V} [\eta_{ki}(0) + \eta_{ki}(4)],\\
B^{+}_{i,V}(s_i = 1) &= \prod_{k\in b(i)\setminus V} [\eta_{ki}(2) + \eta_{ki}(6)],\\
B^{-}_{i,V}(s_i = 0) &= \prod_{k\in b(i)\cap V} [\eta_{ki}(1) + \eta_{ki}(5)],\\
B^{-}_{i,V}(s_i = 1) &= \prod_{k\in b(i)\cap V} [\eta_{ki}(3) + \eta_{ki}(7)],
\end{aligned}$$

where V denotes any subset of residue sites. We define similar quantities for the contact–map messages g:

$$\begin{aligned}
G_{i,V}(s_i = 0) &= \prod_{k\in g(i)\setminus V} [\eta_{ki}(0) + \eta_{ki}(1) + \eta_{ki}(4) + \eta_{ki}(5)], \qquad (3.29)\\
G_{i,V}(s_i = 1) &= \prod_{k\in g(i)\setminus V} [\eta_{ki}(2) + \eta_{ki}(3) + \eta_{ki}(6) + \eta_{ki}(7)],\\
G^{+}_{i,V}(s_i = 1) &= \prod_{k\in g(i)\setminus V} [\eta_{ki}(2) + \eta_{ki}(3) + \eta_{ki}(6)],\\
G^{-}_{i,V}(s_i = 1) &= \prod_{k\in g(i)\cap V} \eta_{ki}(7).
\end{aligned}$$

These products simplify greatly notation in the update formula. Note that the definition formulae (3.28, 3.29) take correctly care of the possibility that some subsets of neighbours may be empty. For example in the first term of (3.27) it may happen that there is no other neighbour of i along backbone than j (*i.e.*, endpoints of sequence). For (3.25) we obtain, after some algebra,

$$\eta_{ij}(0) = \frac{T(0,\bar{s}_i)}{C_{ij}} \left[e^{-\mu} G_i(0) \sum_{m \in b(i) \setminus j} B^-_{i,m}(0) + e^{-2\mu} G_i(0) B^+_{i,j}(0) \right]. \tag{3.30}$$

Now we can write explicit formulae for the other components of message η_{ij} along the backbone. Analogously as before, one finds

$$\eta_{ij}(1) = \eta_{ij}(3) = e^{-\mu}\eta_{ij}(0), \quad \eta_{ij}(2) = \eta_{ij}(0). \tag{3.31}$$

Similarly,

$$\begin{aligned} \eta_{ij}(4) = \frac{T(1,\bar{s}_i)}{C_{ij}} \Big\{ & G_i(1)B_{i,j}(1) + (e^{-2\mu}-1)G^+_i(1)B^+_{i,j}(1) + \\ & + (e^{-\mu}-1)\Big[G^+_i(1) \sum_{m \in b(i)\setminus j} B^-_{i,m}(1) + \\ & + \sum_{m \in g(i)} [G^-_{i,m}(1)G^+_{i,m}(1)]B^+_{i,j}(1)\Big]\Big\}, \\ \eta_{ij}(6) = \eta_{ij}(4). & \end{aligned} \tag{3.32}$$

And finally

$$\begin{aligned} \eta_{ij}(5) = \frac{T(1,\bar{s}_i)}{C_{ij}} \Big\{ & G_i(1)B_{i,j}(1) + (e^{-3\mu}-1)G^+_i(1)B^+_{i,j}(1) + \\ & + (e^{-2\mu}-1)\Big[G^+_i(1) \sum_{m \in b(i)\setminus j} B^-_{i,m}(1) + \\ & + \sum_{m \in g(i)} [G^-_{i,m}(1)G^+_{i,m}(1)]B^+_{i,j}(1)\Big] + \\ & + (e^{-\mu}-1)\Big[\sum_{\{m,n\} \in g(i)} [G^-_{i,\{m,n\}}(1)G^+_{i,\{m,n\}}(1)]B^+_{i,j}(1) + \\ & + \Big(\sum_{m \in g(i)} [G^-_{i,m}(1)G^+_{i,m}(1)]\Big)\Big(\sum_{m \in b(i)\setminus j} B^-_{i,m}(1)\Big)\Big]\Big\}, \\ \eta_{ij}(7) = \eta_{ij}(5). & \end{aligned} \tag{3.33}$$

It is worth noting that among eight possible η's only three are independent. The same is true also for the messages send along the contact map

edges. In this case we only summarize the formulae for the components of the message η_{ij}:

$$\eta_{ij}(0) = T(0,\bar{s_i})G_{i,j}(0)\Big[\, e^{-3\mu}B_i^+(0) + e^{-2\mu}\sum_{m\in b(i)}[B_{i,m}^+(0)B_{i,m}^-(0)] + \\ + e^{-\mu}\sum_{\{m,n\}\in b(i)} B_{i,\{m,n\}}^-(0)\Big], \tag{3.34}$$

$$\eta_{ij}(1) = \eta_{ij}(2) = \eta_{ij}(3) = \eta_{ij}(0).$$

$$\eta_{ij}(4) = \frac{T(1,\bar{s_i})}{C_{ij}}\Big\{\, G_{i,j}(1)B_i(1) + (e^{-3\mu}-1)G_{i,j}^+(1)B_i^+(1) + \tag{3.35}$$
$$+ (e^{-2\mu}-1)\Big[\sum_{m\in g(i)\setminus j}[G_{i,m}^-(1)G_{i,\{m,j\}}^+(1)]B_i^+(1) +$$
$$+ G_{i,j}^+(1)\sum_{m\in b(i)}[B_{i,m}^-(1)B_{i,m}^+(1)]\Big] +$$
$$+ (e^{-\mu}-1)\Big[\sum_{\{m,n\}\in g(i)\setminus j}[G_{i,\{m,n\}}^-(1)G_{i,\{m,n,j\}}^+(1)]B_i^+(1) +$$
$$+ (\sum_{m\in g(i)\setminus j}[G_{i,m}^-(1)G_{i,\{m,j\}}^+(1)])(\sum_{m\in b(i)}[B_{i,m}^-(1)B_{i,m}^+(1)]) +$$
$$+ G_{i,j}^+(1)\sum_{\{m,n\}\in b(i)} B_{i,\{m,n\}}^-(1)\Big]\Big\},$$

$$\eta_{ij}(5) = \eta_{ij}(7) = \eta_{ij}\ (4),$$

$$\eta_{ij}(6) = \frac{T(1,\bar{s_i})}{C_{ij}}\Big\{\, G_{i,j}(1)B_i(1) + (e^{-2\mu}-1)G_{i,j}^+(1)B_i^+(1) +$$
$$+ (e^{-\mu}-1)\Big[\sum_{m\in g(i)\setminus j}[G_{i,m}^-(1)G_{i,\{m,j\}}^+(1)]B_i^+(1) +$$
$$+ G_{i,j}^+(1)\sum_{m\in b(i)}[B_{i,m}^-(1)B_{i,m}^+(1)]\Big]\Big\}.$$

The normalization constant for the conditional probabilities η_{ij} can be found by simple summation of the components

$$C_{ij} = \sum_{t=0}^{7}\eta_{ij}(t). \tag{3.36}$$

As we compare the mechanical properties of the protein with the prediction of the model through temperature factors, we have to find a quantity corresponding to the temperature factors in our model. The temperature factors (B-factors) measure the influence of dynamics and disorder on x-ray scattering from crystals. We concentrate on isotropic temperature factors TF that are

related to the mean–square displacement of the atom $\langle \Delta r^2 \rangle$ through

$$TF = 8\pi^2 \frac{\langle \Delta r^2 \rangle}{3}. \tag{3.37}$$

The closest quantity to which we have access in our model is the dimension of the motion of any given residue. It is closely related to the number of degrees of freedom left on the residue. The average number of DoF in our ensemble can be expressed as (3.24) ($\mu \gg 1$).

$$h_i^{DoF} = \sum_{\{t_{ij}\}} {}' P_i(\{t_{ij}\}) \max\Big\{0; 3 - \sum_{k \in g(i)} s_k s_i x_{ki} - \sum_{k \in b(i)} x_{ki}\Big\}. \tag{3.38}$$

If we use the formula (3.21) to express the configuration probability of the node s_i and its neighbours to be in state given by $\{t_{ji}\}$, $j \in g(i) \cup b(i)$, we obtain:

$$h_i^{DoF} = \sum_{\{t_{ij}\}} {}' \frac{T(s_i, \bar{s}_i)}{C_i} \prod_{k \in g(i) \cup b(i)} \eta_{ki}(t_{ki}) \max\Big\{0; 3 - \sum_{k \in g(i)} s_k s_i x_{ki} - \sum_{k \in b(i)} x_{ki}\Big\} \times$$
$$\exp\Big\{ - \mu \max\Big\{0; 3 - \sum_{k \in g(i)} s_k s_i x_{ki} - \sum_{k \in b(i)} x_{ki}\Big\}\Big\}. \tag{3.39}$$

We can proceed in the calculation of the mean number of DoF on a given residue in the same way as we did for the update formula (3.19). After some lengthy but simple algebra one finds:

$$\begin{aligned} h_i^{DoF} = \frac{T(0, \bar{s}_i)}{C_i} \Big[& 3e^{-3\mu} G_i(0) B_i^+(0) + 2e^{-2\mu} G_i(0) \sum_{k \in b(i)} [B_{i,k}^-(0) B_{i,k}^+(0)] + \\ & + e^{-\mu} G_i(0) \sum_{\{m,n\} \in b(i)} B_{i,\{m,n\}}^-(0) \Big] + \\ + \frac{T(1, \bar{s}_i)}{C_i} \Big[& 3e^{-3\mu} G_i^+(1) B_i^+(1) + 2e^{-2\mu} \Big[\sum_{k \in g(i)} [G_{i,k}^-(1) G_{i,k}^+(1)] B_i^+(1) + \\ & + \sum_{k \in b(i)} [B_{i,k}^-(1) B_{i,k}^+(1)] G_i^+(1) \Big] + \\ & + e^{-\mu} \Big[\sum_{\{m,n\} \in g(i)} [G_{i,\{m,n\}}^-(1) G_{i,\{m,n\}}^+(1)] B_i^+(1) + \\ & + \sum_{k \in g(i)} [G_{i,k}^-(1) G_{i,k}^+(1)] \sum_{k \in b(i)} [B_{i,k}^-(1) B_{i,k}^+(1)] + \\ & + \sum_{\{m,n\} \in b(i)} [B_{i,\{m,n\}}^-(1)] G_i^+(1) \Big] \Big]. \end{aligned} \tag{3.40}$$

where we have calculated products B and G only after reaching self-consistent set of η^*'s. For complete evaluation of the formula (3.40) we have to find

value of the normalization constant C_i. It can be done in the way very similar to the evaluation of the formula for the update of messages η, (3.19). We can separate the calculation of the normalization constant:

$$C_i = \sum_{\{t_{ji}\}, j \in g(i) \cup b(i)} \prime P_i(\{t_{ji}\}), \tag{3.41}$$
$$C_i = C_i^0 + C_i^1,$$

where the two partial terms read:

$$C_i^a = \sum_{\{t_{ji}\}, j \in g(i) \cup b(i), s_i = a} \prime P_i(\{t_{ji}\}), \quad a \in \{0, 1\}. \tag{3.42}$$

For the two terms we can find these formulae:

$$C_i^0 = T(0, \bar{s}_i) \Big[e^{-\mu} G_i(0) \sum_{\{m,n\} \in b(i)} [B^-_{i,\{m,n\}}(0)] + \tag{3.43}$$
$$+ e^{-2\mu} G_i(0) \sum_{k \in b(i)} [B^-_{i,k}(0) B^+_{i,k}(0)] + e^{-3\mu} G_i(0) B_i^+(0) \Big],$$

And finally,

$$\begin{aligned} C_i^1 = T(1, \bar{s}_i) \Big[& G_i(1) B_i(1) + (e^{-3\mu} - 1) G_i^+(1) B_i^+(1) + \\ & + (e^{-2\mu} - 1) \Big[\sum_{k \in g(i)} [G^+_{i,k}(1) G^-_{i,k}(1)] B_i^+(1) + \\ & \quad + G_i^+(1) \sum_{k \in b(i)} [B^+_{i,k}(1) B^-_{i,k}(1)] \Big] + \\ & + (e^{-\mu} - 1) \Big[\sum_{\{m,n\} \in g(i)} [G^+_{i,\{m,n\}}(1) G^-_{i,\{m,n\}}(1)] B_i^+(1) + \\ & \quad + \sum_{k \in g(i)} [G^+_{i,k}(1) G^-_{i,k}(1)] \sum_{k \in b(i)} [B^+_{i,k}(1) B^-_{i,k}(1)] + \\ & \quad + G_i^+(1) \sum_{\{m,n\} \in b(i)} B^-_{i,\{m,n\}}(1) \Big] \Big]. \end{aligned} \tag{3.44}$$

Once knowing the partial contribution of the residues to the total flexibility of the protein we can express the total number of the floppy modes as (3.23). The overall procedure is summarized in the pseudo-code of the Figure 3.8.

- Main routine
  - INPUT: The 3D structure of the protein, the cut-off distance and the number of bonds to be fixed
  - Generate the graphs g and b from the 3D structure.
  - Generate ensemble of subsets $\{s_i\}$ with a given number of bonds.
  - Evaluate flexibility for each subset in the ensemble (subroutine Flexibility) and create the histogram of flexibility. Create ensembles from the subsets that lead to the most rigid, the most flexible structures, and some control set (may be the complete set of generated subsets).
  - OUTPUT: Using the subroutine DoF evaluate the correlation of the DoF profile and the temperature factors in the three ensembles. Plot the histograms of correlation coefficients. Calculate the mean DoF profiles in the three ensembles and compare them with the temperature factors.
- Subroutine Flexibility
  - Given the configuration $\{s_i\}$ calculate the DoF profile for the subset. For computation of the DoF profile use subroutine DoF.
  - OUTPUT: calculate and output total flexibility using (3.24)
- Subroutine DoF
  - INPUT sequence $\{s_i\}$, choose large enough μ
  - For the given subset of active residues iterate in the random order of links (ij) the formulae (3.30, 3.31, 3.32, 3.33, 3.34, 3.35) until self-consistency. Store the self-consistent messages as η^*.
  - Calculate the DoF profile using formula (3.40).

Figure 3.8: The pseudo-code of the procedure that we are using in our calculations.

Chapter 4

Conclusions

In our work the new methods developed within the spin–glass field of statistical physics were applied to the study of two case problems: the random Boolean Satisfiability Problem and the protein flexibility. The methods either allowed observation of new regions of the phase diagram of the problem as in the case of SAT and/or thanks to the speed–up allowed us to study the relation between steric structure of protein and its flexibility.

Results for the SAT Problem

In the complex structure of the phase diagram of the Boolean Satisfiability Problem, there are also regions where the local search heuristics turn out to be inefficient. Indeed, in the studied region close to the SAT/UNSAT transition, $\alpha \in [4.15; 4.39]$, their time complexity grows exponentially with the system size. The local search algorithms are generally trapped at some threshold energy levels that are algorithm specific but that are approximated by the Gardner threshold energy. The threshold energies emerge as a consequence of exponentially many metastable states with energies higher than the ground state energy what is observed through finite complexity at given energy. In the studied region the complexity is growing function of energy until some maximal (threshold) value when it becomes zero.

The difference between the ground state energy and threshold energies becomes well pronounced only when large enough ($N \sim 10^5$) sizes of formulae are reached. Then, as was shown in our study, we are able to see the gap in numerical experiments, and measure its properties.

Conversely to the local heuristics, the Survey Propagation algorithm which explicitly accounts for the clustering of the configuration space is able to approach the ground state solutions and in many cases to find optimal solution of the given instance of the problem. Still, for the intrinsic noise introduced by finite pseudo-temperature, one has to control the procedure with a backtracking procedure, though a very simple one.

It is worth to note that our procedure is perfectly suited for randomly gen-

erated instances of the Satisfiability problem, but it may meet problems when some nontrivial structure is introduced to the formula as it may happen upon coding of some other optimization problem to SAT (as the Graph Colouring, the Scheduling, ...). These problems that are caused by structure–induced correlation of some variables may be solved by coarse–graining of the underlying graph.

Possible extensions of the work lie ready to hand, one can extend the work for other combinatorial optimization problems, and/or improve the optimization algorithm by using variational methods. Another possible application of SP-Y–like algorithms can be found in information theory: one could for example imagine lossy data compression based on Low Density Parity Check schemes with possibility to address exponential number of states by selecting suitable external conditions [82].

Results on the Relation of the Protein Structure and Flexibility

The underlying idea of the second part of the thesis is the expectation that the 3D-structure of a protein determines its function and hence it is this structural properties that are under selective pressure. In many proteins the function is, if not based, strongly influenced by mechanical properties such as flexibility, so we have tried to couple the structure and the flexibility of the protein in a simple physical model.

The coarse grained model is in its simple form described by two terms, structure and flexibility, in terms of the contact energy and the number of constraints obeyed by a residue.

The high level of coarse–graining together with application of message–passing algorithms allows us to study behaviour of the model in a very representative way. The results of our model may be used either directly, or as a method how to estimate model parameters for models with a lower level of course graining and ultimately for models based on microscopical description of the physical system.

To test the scheme we have applied it on a model protein, the HIV protease, and we have obtained prediction for the mobility of single residues consistent with experimental data and with the simulations done by low-level methods (Molecular Dynamics). Still the resulting correlation matrices do not agree very well with Molecular Dynamics. We see the reason for the discrepancy in the simplicity of the model and strength of some approximations that we have used.

In this way we immediately see possible extensions of the model: one could introduce more details to the model, for example by considering also the C_β atoms, by distinguishing several possible bond types and hence introducing a strength of a bond, and finally by using the primary sequence information for estimation of a bond strength.

Physics and Constraint Satisfaction Problems

The two problems that we have discussed in this thesis belong to the group of constraint satisfaction problems over relatively sparse graphs. In the Satisfiability problem the constraint take form of the logical clauses, in the protein flexibility problem they are spatial constraints implied by neighbouring residues. The separation of the cost function (of the number of violated clauses or of the total flexibility of the protein) in many local constraint allowed us to apply to the study of the problems message–passing algorithms.

The constraint satisfaction problems are common to statistical physics and to the field of combinatorial optimization. Naturally the fields reciprocally influence each other and the detailed knowledge of the configuration space provided by spin–glass physics influences choice of proper algorithm to solve the satisfiability problem (the Survey Propagation). On the other side in the latter problem, the configuration space was found to be un-clustered and hence the simpler method of the Belief Propagation has been applied.

Acknowledgments

I want mostly to thank Riccardo Zecchina for his belief in biological applications of physics that has hopefully not diminished during the work on protein topic. And to Cristian Micheletti who helped us to find a field of biological physics where to apply the methods of such a distant field like combinatorial optimization. And there are many other reasons for which I am much obliged to them.

I would like to thank to all my colleagues, starting with Demian Battaglia, who was an irreplaceable partner of mine during the solution of the first part of the work. And I do hope I was also a valuable partner for him. To work together was like Spin Glass music (or Phil?).

At the same place I have to thank to Michele Leone with whom we climbed through the second part of the thesis. There were some falls but the GriGri of criticism helped us to block them and to continue without breaking my head.

I am also indebted to all the colleagues from the Statistical and Biological Physics sector, to Paolo Carloni, Giacomo, Tommaso, now also Osvaldo and many others. As to the colleagues from ISI in Torino, to Andrea who risked to work on gene regulatory networks before there were data available, to Martin and Alfredo.

I wish to thank to whole my family for supporting me although it is over the Alps from Prague to Trieste.

And certainly to all my friends who either even came to check if everything is all right with me or they controlled me on remote :o) And I have even found some new friends here in Trieste, keeping my feet or at least my hands on the limestone rocks of Carso. Thank to Jaro, Juraj and Alessandro for pushing and pulling me up along the nicest routes in Dolomites and Calabroni.

And mostly I would like to thank to Daniela who has become K. instead of C.

Index

Bibliography

Introduction

[1] Parisi, Giorgio, *Statistical Field Theory*, (Addison–Wesley, Wokingham, United Kingdom, 1988)

[2] Cook S.A., The complexity of theorem-proving procedures. In *Proceedings of the Third Annual ACM Symposium on Theory of Computing*, pages 151-158, (1971)

[3] Erdős P. and Rényi A., On Random Graphs I, *Publ. Math. Debrecen* **6** 290 (1959); Bollobás B., *Random Graphs* 2^{nd} Edition, Cambridge University Press (2001)

[4] Metropolis N., Rosenbluth A.W., Rosenbluth M.N., Teller A.H., and Teller E., Equations of State Calculations by Fast Computing Machines, *Journal of Chemical Physics* **21** 1087-1092 (1953)

[5] Newman M.E.J., Barkema, G.T., *Monte Carlo methods in statistical physics*, Oxford Clarendon Press (1999)

MAX-K-SAT

[6] Battaglia D.A., Kolář M., Zecchina R., Minimizing energy below the glass thresholds, *Phys. Rev. E* **70** 036107 (2004)

[7] Battaglia D.A., Kolář M., Zecchina R., From statistical physics methods to algorithms, *Proceedings of Quantum Many-Body theory XII-th workshop, Santa Fe*, to be published in *Int. J. Mod. Phys. B* (2005)

[8] J.-P. Bouchaud, L. F. Cugliandolo, J. Kurchan and M. Mézard, in *Spin Glasses and Random Fields*, edited by A. P. Young (World Scientific 1997)

[9] H. Frauenfelder, P. G. Wolynes, and R. H. Austin, *Rev. Mod. Phys.* **71**, s419–s430 (1999)

[10] V. S. Pande, A. Y. Grosberg, and T. Tanaka, *Rev. Mod. Phys.* **72**, 259–314 (2000)

[11] M. Blatt, S. Wiseman, and E. Domany, *Physical Review Letters* **76**, 3251 (1996)

[12] A. K. Hartmann, H. Rieger, *Optimization Algorithms in Physics* (Wiley-VCH, Berlin, 2001)

[13] A. K. Hartmann, H. Rieger, *New and Advanced Optimization Algorithms in Physics and Computational Science* (Wiley-VCH, Berlin, 2004)

[14] S. Kirkpatrick, C. D. Gelatt, M. P. Vecchi, Optimization by Simulated Annealing, [http://citeseer.ist.psu.edu/kirkpatrick83optimization.html] *Science* **220**, 671–680 (1983)

[15] C. H. Papadimitriou, K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity* (Prentice-Hall, Englewood Cliffs, NJ, 1982)

[16] Special Issue on *NP-hardness and Phase transitions*, edited by O. Dubois, R. Monasson, B. Selman and R. Zecchina, *Theor. Comp. Sci.* **265**, Issue: 1-2 (2001)

[17] T. Hogg, B.A. Huberman, C. Williams, C. (eds), *Artificial Intelligence* **81** I & II (1996)

[18] H. Nishimori, *Statistical Physics of Spin Glasses and Information Processing* (Oxford University Press, 2001)

[19] D. A. Spielman, *Lecture Notes in Computer Science* **1279**, pp. 67-84 (1997)

[20] T. Richardson and R. Urbanke, An introduction to the analysis of iterative coding systems, in *Codes, Systems, and Graphical Models*, edited by B. Marcus and J. Rosenthal (Springer, New York, 2001)

[21] N. Sourlas, *Nature* **339**, 693 (1989)

[22] N. Ajtai, *Electronic Colloquium on Computational Complexity (ECCC)* **7**, 3 (1996)

[23] R. Monasson, R. Zecchina, S. Kirkpatrick, B. Selman, and L. Troyanski, *Nature* **400**, 133 (1999)

[24] Mézard M., Parisi G., Zecchina R., Analytic and Algorithmic Solution of Random Satisfiability Problems, *Science* **297** 812 (2002) (doi:10.1126/science.1073287)

[25] S. Cocco, R. Monasson, A. Montanari, and G. Semerjian, Approximate analysis of search algorithms with "physical" methods, *preprint cs.CC/0302003* (2003)

[26] Mézard M., Zecchina R., Random K-satisfiability problem: From an analytic solution to an efficient algorithm, *Phys. Rev. E* **66**, 056126 (2002)

[27] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, *IEEE Trans. Inf. Theory* **47**, 498 (2002)

[28] A. Braunstein, M. Mézard, R. Zecchina, Survey propagation: an algorithm for satisfiability, preprint 2002, to appear in *Random Structures and Algorithms*, *cs.CC/0212002*

[29] A. Montanari, F. Ricci-Tersenghi, On the cooling-schedule dependence of the dynamics of mean-field glasses *preprint, cond-mat/0401649* (2004)

[30] R. Motwani, P. Raghavan, *Randomized Algorithms* (Cambridge University Press, Cambridge, 2000)

[31] G. Semerjian and R. Monasson, Relaxation and Metastability in the Random WalkSAT search procedure, *cond-mat/0301272, preprint* (2003)

[32] W. Barthel, A. K. Hartmann, and M. Weigt, Solving satisfiability problems by fluctuations: An approximate description of the dynamics of stochastic local search algorithms, *cond-mat/0301271, preprint* (2003)

[33] G. Parisi, A backtracking survey propagation algorithm for K-satisfiability, *preprint cond-mat/0308510* (2003)

[34] M. R. Garey and D. S. Johnson, *Computers and intractability* (Freeman, New York, 1979)

[35] S. Kirkpatrick, B. Selman, *Science* **264**, 1297 (1994)

[36] E. Friedgut, *Journal of the A.M.S.* **12**, 1017 (1999)

[37] O. Dubois, Y. Boufkhad, and J. Mandler, Typical random 3-SAT formulae and the satisfiability threshold, in *Proc. 11th ACM-SIAM Symp. on Discrete Algorithms*, 124 (San Francisco, CA, 2000); A. Kaporis, L. Kirousis, and E. Lalas, The probabilistic analysis of a greedy satisfiability algorithm, in *Proceedings of the 4th European Symposium on Algorithms* (ESA 2002), to appear in series: *Lecture Notes in Computer Science*, Springer

[38] F. Guerra, *Comm. Math. Phys.* **233**, 1 (2003); S. Franz, M. Leone *J. Stat. Phys.* **111**, 535 (2003)

[39] J. Franco, *Theoretical Computer Science* **265**, 147 (2001); D. Achlioptas, G. Sorkin, *41st Annu. Symp. of Foundations of Computer Science, IEEE Computer Soc. Press*, 590 (Los Alamitos, CA, 2000)

[40] D. Achlioptas, C. Moore, Random k-SAT: Two Moments Suffice to Cross a Sharp Threshold, *preprint* (2002)

[41] D. Achlioptas, U. Noar, Y. Peres, On the Fraction of Satisfiable Clauses in Typical Formulas, *preprint, extended abstract FOCS'03*, pp. 362-370

[42] U. Schöning, *Algorithmica* **32**, 615-623 (2002)

[43] M. Alekhnovich, E. Ben-Sasson, Analysis of the Random Walk Algorithm on Random 3-CNFs, *preprint* (2002)

[44] A.J. Parkes, *Lecture Notes in Computer Science* **2470**, 708 (2002)

[45] H. Karlo, U. Zwick, In *Proc. of 38th FOCS*, 406–415 (1997)

[46] S.Mertens, M. Mézard, R. Zecchina, Threshold values of Random K-SAT from the cavity method, preprint, *cs.CC/0309020* (2003)

[47] G. Biroli and R. Monasson, and M. Weigt, *Eur. Phys. J.* **B 14**, 551 (2000)

[48] G. Parisi, Some remarks on the survey decimation algorithm for K-satisfiability, *preprint cs.CC/0301015* (2003).

[49] Annibale A., Gualdi G. and Cavagna A, Coexistence of supersymmetric and supersymmetry–breaking states in spherical spin-glasses, *J. Phys. A: Math. Gen.* 37 (2004) 11311–11320

[50] A. Montanari, G. Parisi, and F. Ricci-Tersenghi, *J. Phys. A* 37, 2073 (2004)

[51] A.V. Lopatin, L.B. Ioffe, *Phys. Rev.* **B 66**, 174202 (2002)

[52] Seitz S., Alava M., ana Orponen P., Focused Local Search for Random 3-Satisfiability, *J. Stat. Mech.* (2005) P06006, [http://www.citebase.org/cgi-bin/citations?id=oai:arXiv.org:cond-mat/0501707]

[53] A.Z. Broder, A.M. Frieze, E. Upfal, in *Proc. 4th Annual ACM-SIAM Symp. on Discrete Algorithms*, 322 (1993)

[54] B. Selman, H. Kautz and B. Cohen, *Proc. AAAI-94*, Seattle, WA, 337-343 (1994)

[55] S. Seitz, P. Orponen: An efficient local search method for random 3-satisfiability, in *Proc. LICS'03 Workshop on Typical Case Complexity and Phase Transitions* (Ottawa, Canada, June 2003); Electronic Notes in Discrete Mathematics Vol. 16. (Elsevier, Amsterdam, 2003)

[56] download site: `www.ictp.trieste.it/~zecchina/SP`

Protein Flexibility

[57] Yedidia J.S., Freeman W.T, and Weiss Y., Understanding Belief Propagation and its Generalizations, *Mitsubishi electric research laboratories* (http://www.merl.com) technical report TR-2001-22 (2002)

[58] Mézard M., Parisi G., The cavity method at zero temperature, *J. Stat. Phys.* **111** 1 (2003)

[59] Jacobs D.J. and Hendrickson B., An Algorithm for Two-Dimensional Rigidity Percolation: The Pebble Game, *Journal of Computational Physics* **137** 346–365 (1997)

[60] Special Issue on *NP-hardness and Phase transitions*, edited by O. Dubois, R. Monasson, B. Selman and R. Zecchina, *Theor. Comp. Sci.* **265**, Issue: 1-2 (2001)

[61] Murphy K.P., Weiss Y., and Jordan M.I., Loopy Belief Propagation for Approximate Inference An Empirical Study. In *Proceedings of Uncertainty in AI* 467-475 (1999) (http://citeseer.ist.psu.edu/murphy99loopy.html)

[62] Weiss Y. and Freeman W.T., Correctness of Belief Propagation in Gaussian Graphical Models of Arbitrary Topology, *Neural Computation* **13** 2173–2200 (2001) (citeseer.ist.psu.edu/article/weiss99correctness.html)

[63] Rutenber, E., Fauman, E.B., Keenan, R.J., Fong, S., Furth, P.S., Ortiz de Montellano, P.R., Meng, E., Kuntz, I.D., DeCamp, D.L., Salto, R., et al.: Structure of a non-peptide inhibitor complexed with HIV-1 protease. Developing a cycle of structure-based drug design. *J Biol Chem* **268** 15343 (1993)

[64] Laman, G., On graphs and rigidity of plane skeletal structures, *J. Eng. Math.* **4**331-340 (1970)

[65] Hespenheide, B., Rader A.J., Thorpe M.F., Kuhn L.A., Identifying protein folding cores from the evolution of flexible regions during unfolding, *Journal of Molecular Graphics and Modelling* **21** 195-207 (2002); Rader A.J., Hespenheide B.M., Kuhn L.A. and Thorpe M.F., Protein unfolding: rigidity lost, *PNAS USA*, **99** 3540-5 (2002); Jacobs D.J., Rader A.J., Kuhn L.A. and Thorpe M.F., Protein flexibility predictions using graph theory, *Proteins* **44** 150-65 (2001); Jacobs D.J., Kuhn L.A. and Thorpe M.F., Flexible and rigid regions in proteins, *in* Rigidity theory and applications, M.F. Thorpe and P.M. Duxbury, Editors (1999) Academic/Kluwer. p. 357-384; Rader A.J. and Bahar I., Folding core predictions from network models of proteins, *Polymer* 655-668 (2004)

[66] Barré J.,Bishop A.R.,Lookman T., and Saxena A., The Cavity Method for the Rigidity Transition, *Journal of Statistical Physics* Vol. **118**, Nos. 5/6 (2005).

[67] Zwanzig, R., Simple model of protein folding kinetics, 1995 *PNAS USA* **92** 9801–9804

[68] Shakhnovich E.I., Proteins with selected sequences fold into unique native conformation, 1994 *Phys. Rev. Lett.* **72** 3907

[69] Muñoz V., Eaton W.A., A simple model for calculating the kinetics of protein folding from three-dimensional structures, 1999 *PNAS USA* **96** 11311–11316

[70] Galzitskaya O.V., Finkelstein A.V., A theoretical search for folding/unfolding nuclei in three-dimensional protein structures, 1999 *PNAS USA* **96** 11299–11304

[71] Alm E., Baker D., Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures, 1999 *PNAS USA* **96** 11305–11310

[72] Muñoz V., Henry E.R., Hofrichter J., Eaton W.A., A statistical mechanical model for β-hairpin kinetics, 1998 *PNAS USA* **95** 5872–5879

[73] Schellman J., The factors affecting the stability of hydrogen-bonded polypeptide structures in solution, 1958 *J. Phys. Chem.* **62** 1485–1494

[74] Gō N. and Scheraga H.A., On the Use of Classical statistical mechanics in the treatment of polymer chain conformation, 1976 *Macromolecules* **9** Number 4 July-August

[75] Miyazawa S. and Jernigan R.L, Residue–Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term, for Simulation and Threading, *J. Mol. Biol.* **256** 623–644 (1996)

[76] D'Aquino J.A., Gomez J., Hilser V.J., Lee K.H., Amzel L.M., and Freire E., The Magnitude of the Backbone Conformational Entropy Change in Protein Folding, *PROTEINS: Structure, Function, and Genetics* **25** 143-156 (1996)

[77] Micheletti C., Carloni P., and Maritan A., Accurate and Efficient Description of Protein Vibrational Dynamics: Comparing Molecular Dynamics and Gaussian Models, *PROTEINS: Structure, Function, and Bioinformatics* **55** 635–645 (2004)

[78] Cecconi F., Micheletti C., Carloni P., and Maritan A., Molecular Dynamics Studies on HIV-1 Protease Drug Resistance and Folding Pathways, *PROTEINS: Structure, Function, and Genetics* **43** 365–372 (2001)

[79] Micheletti C., Cecconi F., Flammini A., and Maritan A., Crucial stages of protein folding through a solvable model: Predicting target sites for enzyme-inhibiting drugs, *Protein Science* **11** 1878–1887 (2002)

[80] Bahar I., Burak Erman B., Jernigan R.L., Atilgan A.R., and Covell D.G., Collective Motions in HIV-1 Reverse Transcriptase: Examination of Flexibility and Enzyme Function, *J. Mol. Biol.* **285** 1023-1037 (1999)

[81] Halle, B., Flexibility and packing in proteins. *PNAS USA* **99** 1274–1279 (2002)

Conclusions

[82] Ciliberti S., Mézard M., Zecchina R., Lossy Data Compression with Random Gates. *Phys. Rev. Letters* **95** 038701 (2005)