ments show a high correlation between the number of basis vectors of a lattice and its ability to represent a protein backbone[21,22]. When deciding on a lattice model, one must always consider the trade-off between the reduction of the conformational space and the quality of the structure representation. Therefore, in section *Lattice experiments* we evaluate four different lattices of various complexity: The SCC lattice, the FCC lattice and two *high coordination* (HC) lattices with 54 and 390 basis vectors, respectively.

A high coordination lattice has an underlying cubic lattice with unit length less than $3.8/N$ Å for some integer $N > 1$. Cubic lattice points are connected in the high coordination lattice if their Euclidean distance is between $3.8 \pm \varepsilon$ for some $\varepsilon > 0$. The high coordination lattices used here are named HC4 and HC8 corresponding to their N value (4 and 8). The ε value is 0.2 for all HC lattices. Figure 3 shows an illustration of a 2D high coordination lattice with $N = 3$ and $\varepsilon = 0.4$. High coordination lattices have previously been used for protein structure prediction[23,24]. Note that the SCC and FCC lattices both have the excluded volume property, meaning that atoms at two different lattice points will never collide. This property does not necessarily hold for high coordination lattices, and collisions must therefore be detected explicitly.

Heuristics

We apply two iterative search heuristics for minimization of the HSE energy. One of them is the tabu search *metaheuristic* proposed by F. Glover in 1989[25,26]. A metaheuristic is a general framework that can be specialized to solve various optimization problems. For many problems in Operations Research (OR), tabu search is the metaheuristic of choice. However, for protein structure prediction, tabu search has only been given a modest amount of attention[14-16].

In Algorithm 1 and 2 (Figures 5 and 6) the pseudo code for tabu search is shown. TS is basically a local improvement heuristic where the best structure in a neighbourhood is repeatedly selected. However, memory is used to prevent cycling in local minima. A previous TS implementation [16] inserts visited structures into a *tabu list* and only consider new structures if they are not in the tabu list. We have found that extending the tabu definition improves the performance considerably. Here, we still keep a list of previously visited structures in a so-called *explicit tabu list*. Each structure in the explicit tabu list defines a set of *implicit tabu structures*. Given a structure E in the explicit tabu list, a structure I is said to be implicit tabu if the distance-RMSD (dRMSD) between E and I is less than ε and the energy of I is greater than or equal to the energy of E. The adjustable parameter ε is called the *tabu difference*. Figure 4 illustrates a sequence of visited structures (black points) in a solution space. Only the visited structures are inserted in the explicit tabu list. The additional green and red points correspond to structures within ε dRMSD of the explicit tabu structures. Green points are structures with lower energy and red points are structures with higher energy than the explicit tabu structure. When choosing a new solution in the neighbourhood three things can happen, *a)* A solution is more than ε dRMSD away from all explicit tabu structure. *b)* the solution is within ε dRMSD, and the energy is *lower* than the explicit tabu structure, *c)* the solution is within ε dRMSD, and the energy is *higher* than the explicit tabu structure. Structures that comply with case *c* are said to be *implicit tabu* and cannot be visited. Note that when $\varepsilon = 0$ the search heuristic works as a regular TS heuristic since only visited structures become tabu. The use of implicit tabu structures is new in the context of protein structure prediction. However, in TS implementations for OR problems it is a common technique to make features of a solution tabu, such that regions of the search space become tabu.

We have also applied standard Monte Carlo simulation (MCS) for minimizing the HSE energy. MCS heuristics are stochastic and therefore differ from TS by being nondeterministic. An MCS iteration consists of randomly choosing a protein conformation in the neighbourhood of a current conformation. For a fixed temperature T, the new protein conformation is accepted with the probability

$$p = e^{-\Delta E/T},$$

where ΔE is the difference between the energy of the current conformation and the new conformation. A protein conformation is modelled as a list of N vectors, where N is the number of C_α atoms of the protein. The neighbourhood of both MCS and TS consists of conformations resulting from changes of one, two or three consecutive indices. A single index change results in a new structure where one part of the structure is fixed and the other part is translated. Two or three indices are changed locally such that the parts of the structure before and after the changing indices are fixed. All local index changes between two lattice points can be stored in a table to speed up the computation time significantly.

Lattice experiments

Here, we evaluate TS and MCS on lattices of different complexity. The purpose of the experiments in this section is to tune the parameters (lattice type, tabu difference, temperature). In the next section we fix the parameters to their optimal values found here and compare the HSE and CN measures on different proteins. For each lattice, the heuristics are initialized with 20 random conformations using different parameter values. The variable parameter of MCS is the temperature and the variable parameter of TS is the