JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Dynamics in Sequence Space for RNA Secondary Structure Design

Marco C. Matthies,*[†] Stefan Bienert,[†,‡] and Andrew E. Torda*[†]

[†]Centre for Bioinformatics, University of Hamburg, Bundesstr. 43, 20146 Hamburg, Germany

[‡]Biozentrum, University of Basel, Klingelbergstr. 50/70, 4056 Basel, Switzerland

ABSTRACT: We have implemented a method for the design of RNA sequences that should fold to arbitrary secondary structures. A popular energy model allows one to take the derivative with respect to composition, which can then be interpreted as a force and used for Newtonian dynamics in sequence space. Combined with a negative design term, one can rapidly sample sequences which are compatible with a desired secondary structure via simulated annealing. Results for 360 structures were compared with those from another nucleic acid design program using measures such as the probability of the target structure and an ensemble-weighted distance to the target structure.

INTRODUCTION

In the problem of polymer sequence design, one has a target structure which is fixed, but a sequence which has to be optimized with respect to some design criterion, such as the similarity of the designed molecule to the target structure. Here, we suggest a method which may be best suited to RNA design or how to find an appropriate sequence given an RNA secondary structure. It is based on completely rigorous Newtonian dynamics simulations, but in an entirely nonphysical four-dimensional sequence space. Furthermore, we are only interested in the most general case where there are no restrictions on composition, no biases toward a native sequence, and no particular starting sequence.

RNA is a friendly polymer to design, as the nearest-neighbor model for RNA secondary structures combined with dynamic programming allows efficient and, considering the simplicity of the model, surprisingly effective structure prediction as well as calculation of thermodynamic properties.[1] Working with this model allows us to predict the structures of designed sequences and compare them to the desired target structure.

We consider RNA sequence optimization, but in many ways the problem is identical to protein design. One has some score function and, for n residues, a search space of 4^n or 20^n possible discrete solutions. From this point of view, one should be able to use many of the same techniques from the protein design literature.[2,3]For this kind of discrete problem, the natural approach would be classic Monte Carlo simulation[4] where trial moves involve changing residue type or rotamer.[5−7] One would expect this to be combined with simulated annealing[8] and perhaps some kind of bias and more elaborate acceptance criterion so as to maintain detailed balance.[9−11] In the specific area of RNA design, one may use a more greedy acceptance criterion but gain speed by using a cleverly selected initial sequence or decomposing the secondary structure graph into, hopefully, independently designable pieces.[12−15] Continuing to see this as a search over discrete variables, one could also prune away most impossible solutions[16−18] and perhaps leave a set which can be systematically searched.[19]

All of these search methods reflect the discrete nature of the problem. One wants real bases or amino acids in the answer. At the same time, there is no reason why a search method should not visit nonphysical configurations en route to the final answer. This philosophy leads to methods such as self-consistent mean fields, in which a site may have a probability of being in different states (residue/rotamer types) simultaneously. One only needs to find whole residue types as the system converges.[20−23] Continuing in this vein, one can have a mixed Hamiltonian where the energy is given by

$$E(\lambda) = \lambda E_1 + (1 - \lambda)E_2 \tag{1}$$

where E_1 and E_2 are energies associated with some states 1 and 2 and λ may be a dynamic variable much like Cartesian coordinates.[24] Taking the derivative of $E(\lambda)$ with respect to λ leads to λ-dynamics as used in free energy calculations.[25] Our work has much in common with these extended Hamiltonian methods, but with a different space and with the goal of Newtonian dynamics in this space.

Figure 1 outlines a system with an example RNA fragment drawn in the conventional two-dimensional manner. Each of the 13 particles represents a base which is fixed in the physical space. The wavy lines, however, represent the nonphysical dimensions which one could call λ coordinates or sequence space. Each particle's λ coordinate reflects how much of each kind of RNA base it consists of (A, C, G, or U). The energy $E(\lambda)$ or score is calculated by some model and the force on particle i given by $-\partial E/\partial \lambda_i$. For example, a particle mostly in the "A" direction will tend to send any base-paired neighbor toward the "U" direction. It is then a simple step to assign every particle a unit mass and run a Newtonian dynamics simulation. In the nearest-neighbor energy model used here,[1] the forces could be seen as acting on the coefficients in front of table-lookup terms.

In this approach, there are several restrictions to be enforced. The total composition at any site should be near 1, and each individual component is in a range from 0 to 1. It does not matter if the system temporarily visits impossible config-

Special Issue: Wilfred F. van Gunsteren Festschrift

Received: March 31, 2012
Published: August 16, 2012

dx.doi.org/10.1021/ct300267j | J. Chem. Theory Comput. 2012, 8, 3663−3670

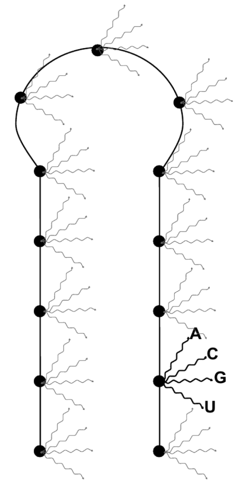

Figure 1. Fictitious dimensions and real coordinates. Nonphysical dimensions are marked by wavy lines. Base types for the nonphysical dimensions are marked at one site only.

urations, such as negative composition, so only soft restraints were used.

As described, the method works as a dynamic system but immediately falls prey to a problem well-known in protein and RNA design. If there are no restrictions, the system will maximize the number of favorable interactions. In many protein score functions, this leads to an overwhelmingly hydrophobic sequence.[26,27] In an RNA model, the equivalent would be a sequence of mostly GC pairs. There are many solutions of varying elegance to the problem. One could limit the search to shuffling an existing sequence or staying close to the composition of a starting sequence[28−30] or even attempt to calculate the change in structure and penalize the system if it drifts away from the target structure.[31] Another approach used in protein sequence design is the use of a foldability criterion, in which the unfolded state is characterized by mean-field theory.[32] In RNA sequences, one may penalize the system if its predicted base pairing differs too much from the desired structure.[12] In this work, we treat this problem of negative design explicitly via the energy/score function.[33] Regardless of the exact score function, each site has interaction neighbors with whom it should interact. It also has sites with whom it should not form favorable interactions. For example, a base whose λ points mostly in the "A" direction may drag a neighboring residue toward "U", but one would like it to push other bases toward A, C, or G. This is easily done by treating these "negative" interactions via the same energy/score function, but with a different sign and an appropriate weighting. In this scheme, the system as a whole will search for some arrangement in λ (sequence) space which produces the desired interactions but minimizes unwanted attraction between particles.

Nucleotide sequence design differs from typical λ-dynamics in some practical respects. If one is dealing with two or a half a dozen potential ligands, searching along λ degrees of freedom is not a major problem. One can even apply a bias to specific λ dimensions to change the sampling behavior.[34] In nucleotide design, there are 4^n possible sequences, so one has no overview of the space. The negative-design term that we apply is a distortion of the energy landscape but definitely not directed toward any specific composition.

METHODS

Nearest-Neighbor Model Energies for Sequence Compositions. The nearest-neighbor model[1] assumes a decomposition of an un-pseudoknotted RNA secondary structure $\omega = \{(i,j) | i$ and j are base paired$\}$ into different loop types (see Figure 2) and approximates the free energy difference between the unfolded state and a given secondary structure ω as a sum of loop contributions:

$$\Delta G_{\text{fold}}(s, \omega) = \sum_{L \in \text{Loops}(\omega)} \Delta\Delta G_L(s_L) \tag{2}$$

where s_L is the sequence of nucleotides inside loop L and $\Delta\Delta G_L(s_L)$ is the loop's free energy contribution according to the nearest-neighbor model.

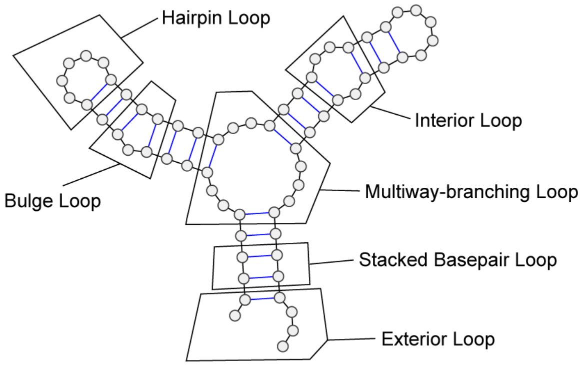

Figure 2. An RNA secondary structure with different loop types shown. Nucleotides are drawn as circles. Base pairs are indicated by blue lines, and nucleotides and base pairs belonging to the same loop are indicated by polygonal outlines. For clarity, only one example of each loop type is shown. Drawn with VARNA.[35]

In this work, we deal with sequence compositions given as a set of composition vectors $\{\lambda\}$. The fraction of base type $\alpha \in \{\text{A,C,G,U}\}$ present at position i is given by $\lambda_{i,\alpha}$. The sequence compositions $\{\lambda\}$ then induce a probability distribution $P_{\{\lambda\}}(s)$ on the space of sequences, and the probability of a given (possibly discontiguous) subsequence s of nucleotides is given by

$$P_{\{\lambda\}}(s) = \prod_{i \in I(s)} \lambda_{i,s[i]} \tag{3}$$

where $I(s)$ is the set of indices defining the subsequence and $s[i]$ is the base type at position i.

Using this probability distribution, we can extend the free energy change given by the nearest-neighbor model to sequence compositions:

$$\Delta G_{\text{fold}}(\{\lambda\}, \omega) = \sum_{L \in \text{Loops}(\omega)} \sum_{s_L} P_{\{\lambda\}}(s_L) \Delta\Delta G_L(s_L) \tag{4}$$

where we assume $\Delta\Delta G_L(s_L) = 0$ if the nucleotide sequence s_L is incompatible with the loop L, as the nearest-neighbor model only deals with the Watson−Crick base pairs GC, AU, and the "wobble" base pair GU.

The sum over all sequences weighted by their probability is just the expected loop free energy contribution of all possible sequences in the loop. In the nearest-neighbor model, the free energy contribution of a loop is only sequence specific at a fixed number of positions independent of loop size: the base pair defining the loop and, depending on the specific loop type under consideration, a finite number of nominally unpaired bases inside the loop (interior and bulge loops, "dangling" bases in multiway-branching loops, hairpin tetraloops). Therefore, the second sum over all subsequences s_L in eq 4 in fact only has to be taken over all sequences where the sequence-specific positions of loop L are varied and the number of sequences to be summed over is a fixed number dependent on the loop type but independent of loop size.

Negative Design Term. The negative design penalizes all possible base-pairing interactions that are not present in the target structure. As there will be on the order of N unwanted interactions for each base, we normalize the total negative design score contribution by a factor of $1/N$.

$$V_{\mathrm{neg}}(\{\lambda\}, \omega^*) = \frac{k_{\mathrm{neg}}}{N} \sum_{(i,j)\notin\omega^*, i<j} \lambda_i^T E_{\mathrm{neg}} \lambda_j \tag{5}$$

Here, ω^* is the target secondary structure, again defined as a set of base pairs. The unwanted base-pairing interactions are modeled in the simplest possible way as a 4×4 matrix E_{neg} that contains the penalty for each possible base pair type combination, so that the product $\lambda_i^T E_{\mathrm{neg}} \lambda_j$ is the expected score contribution of positions i and j.

If the negative design score is calculated from this formula, it will run in quadratic time. Fortunately, it can be rewritten to run in linear time by first calculating the negative design score as if all interactions were unwanted and then subtracting the contribution of the interactions that are present in the target structure:

$$V_{\mathrm{neg}} = \frac{k_{\mathrm{neg}}}{2N}\left(\left(\sum_{i=1}^{N} \lambda_i\right)^T E_{\mathrm{neg}} \left(\sum_{i=1}^{N} \lambda_i\right) - \sum_{i=1}^{N} \lambda_i^T E_{\mathrm{neg}} \lambda_i - 2 \sum_{(i,j)\in\omega} \lambda_i^T E_{\mathrm{neg}} \lambda_j\right) \tag{6}$$

Sequence Heterogeneity Term. The sequence heterogeneity term measures the difference in sequence composition inside a window of width w to the left and right of each position. The difference in sequence composition is expressed here as the cosine of the angle between two sequence composition vectors λ_i and λ_j:

$$V_{\mathrm{het}} = k_{\mathrm{het}} \sum_i \sum_{j=i-w}^{i+w} (1 - \delta_{ij}) \frac{\lambda_i^T \lambda_j}{\|\lambda_i\| \, \|\lambda_j\|} \tag{7}$$

where δ_{ij} is the Kronecker delta.

Composition Constraint Terms. The compositions are restrained to be between 0 and 1, and the sum of compositions at each position is restrained to unity. Both restraints are implemented by a simple harmonic term.

$$V_{\mathrm{cmp}} = \frac{k_{\mathrm{cmp}}}{2} \sum_i \left(\sum_\alpha \lambda_{i\alpha} - 1\right)^2 + \frac{k_{\mathrm{cmp}}}{2} \sum_i \sum_\alpha \sigma(\lambda_{i\alpha}) \tag{8}$$

where

$$\sigma(x) = \begin{cases} 0 & x \in [0, 1] \\ x^2 & x < 0 \\ (x-1)^2 & x > 1 \end{cases} \tag{9}$$

Design Scoring Function. The overall design scoring function is

$$V(\{\lambda\}, \omega^*) = \Delta G_{\mathrm{fold}}(\{\lambda\}, \omega^*) + V_{\mathrm{neg}}(\{\lambda\}, \omega^*) + V_{\mathrm{het}}(\{\lambda\}) + V_{\mathrm{cmp}}(\{\lambda\}) \tag{10}$$

and can be calculated in linear time.

In order to make scores between structures of different sizes comparable, we will sometimes use a normalized design score, which is the design score divided by sequence length.

Sequence Design via Dynamical Simulated Annealing in Sequence Space. The scoring function is differentiable with respect to sequence composition. By interpreting the score as an energy and its gradient with respect to sequence composition as a force, we obtain a Hamiltonian system where the dynamics take place in sequence space.

The equations of motion were integrated numerically using the leapfrog integrator.[36] The sequence composition velocities were restrained by instantaneously rescaling velocities. After a short equilibration at high temperatures, the system was linearly cooled to a temperature of zero.

At some positions in the RNA sequence, the ratio between the free energy of folding as predicted by the nearest-neighbor model and the negative design term might be difficult to get right. If the negative design term is too strong, it can mean that bases that are nominally base-paired in the target structure can be assigned base types that are incompatible according to the nearest-neighbor model. Bases that are supposed to be base-paired but do not have compatible base types were fixed after an optimization run by assigning a random base compatible with base-pairing.

In a typical molecular dynamics simulation, the net center of mass movement is removed to avoid the "flying ice cube" effect.[37] Due to the lack of translational invariance of our scoring function, it follows from Noether's theorem[38] that our Hamiltonian dynamics in sequence space do not preserve overall momentum in λ space (sequence space). We therefore do not remove net center of mass movement of the λ variables, which would be equivalent to enforcing a constant average sequence composition.

This design method is referred to below as "DSS-Opt" (Dynamics in Sequence Space Optimization; see *Note*).

Evaluating Designed Sequences. Obviously, a designed sequence should fold to the target structure, but within the RNA framework, one can use more sophisticated tests. First, one can always predict the secondary structure within cubic time[1] and check if the target structure really is the structure of minimum energy by calculating the base pair distance between the target and minimum free energy structures:

$$d_{\mathrm{bp}}(\omega_1, \omega_2) = |\omega_1 \Delta \omega_2| = |\omega_1| + |\omega_2| - 2|\omega_1 \cap \omega_2| \tag{11}$$

The same dynamic programming approach used for minimum energy structure prediction can then be used to calculate the partition function over allowed secondary structures[39] and thus the probability of the ground state structure

$$P_s(\omega^*) = \frac{1}{Z(s)}\, e^{-\Delta G(s,\omega^*)/RT} \qquad (12)$$

Values should be as close to 1 as possible.

Given that one can estimate the probability of alternative conformations, one can estimate the probability of wrong base pairs. This leads to a measure known as the ensemble defect.[40] It has the advantage that errors are weighted with their Boltzmann probabilities from the ensemble of allowed structures. The probability p_{ij} of base pair (i, j) existing is estimated from the ensemble.[39] One then defines a function $\omega^*(i, j)$ which returns 1 if bases i and j are in the set of base pairs in the target structure and 0 otherwise.

$$d_n(\omega^*) = \sum_{i=1}^{N} (1 - \omega^*(i, j))\left(\sum_{j=1}^{N} p_{ij}\right) + \omega^*(i, j)(1 - p_{ij}) \qquad (13)$$

The ensemble defect can be normalized by dividing by the length of the sequence, thereby making ensemble defects from structures with different sizes comparable.

Force Constants in Scoring Function. Force constants for eq 10 in reduced units were chosen as $k_{\text{neg}} = 1$ for the negative design term, $k_{\text{het}} = 10$, $w = 3$ for the sequence heterogeneity term, and $k_{\text{cmp}} = 50\,000$ for the sequence composition term. Total simulation time was set to 50 time units with a time step for the integrator of 0.0015. The starting sequence temperature was set to 40, and after 5 time units the temperature was linearly cooled until it reached 0.1 at the end of an optimization run.

Test Sets and Testing Procedure. We used the same test sets that were used in the evaluation of the sequence design method contained in NUPACK.[15] The test sets consist of target structures of sizes 100, 200, 400, 800, 1600, and 3200 nucleotides and are divided into two classes: "random" structures are the minimum free energy structures of random sequences as predicted by the nearest-neighbor model, whereas "engineered" structures were selected to more closely resemble man-made secondary structures that typically are more duplex-rich and have fewer multiloops than the structures from the random set. There is a test set for each combination of target size (100, 200, 400, 800, 1600, 3200) and test set type (random or engineered), resulting in 12 test sets with 30 target structures each, for a total of 360 target structures.

For each tested method and each target structure, we collected all sequences designed within one hour and evaluated them according to the criteria explained above. Evaluation of designed sequences was done with our own implementation of the ensemble defect based on the partition function calculation as implemented in the Vienna-RNA library,[12] version 1.8.5. The default nearest-neighbor parameters were used. The temperature was set to 37 °C, and the dangles option was set to "all".

NUPACK needs to be given a stop criterion for the desired normalized ensemble defect, which we left at the default value of 1%. We used NUPACK version 3.0, the "rna1999" parameter set, a temperature of 37 °C, and the dangles option set to "all".

Results were also calculated for random sequences, purely to provide a statistical background. Sequences were generated by randomly assigning base types to unpaired positions, whereas for base paired positions, one of the permissible base pairs within the nearest-neighbor energy model was randomly chosen. These sequences contain Watson−Crick base pairing but are otherwise random. In the Results section, they are referred to as "random compatible".

RESULTS

Comparison of Design Score with Ensemble Defect. Before considering the problem of design, one can assess the score function since it includes the free energy approximation, as well as a negative design term. Ideally, the more negative the value from eq 10, the lower should be the probability of wrong base pairs. One can then plot the ensemble defect (eq 13) as a function of the score function. Figure 3 shows this for 10 000 sequences ranging from 100 to 3200 bases (taken from the test sets). Values are normalized by sequence length, and the sequences are taken from the "random compatible" set; i.e., bases have been chosen so that a canonical Watson−Crick base pair is present at every base pair. The plot shows that the ensemble defect is highly correlated with the score, but the relationship is definitely not perfect. Looking at the left-hand side of the plot with the best scoring sequences, it would be fair to say that even structures with the best possible design score will have some errors in terms of the ensemble defect measure.

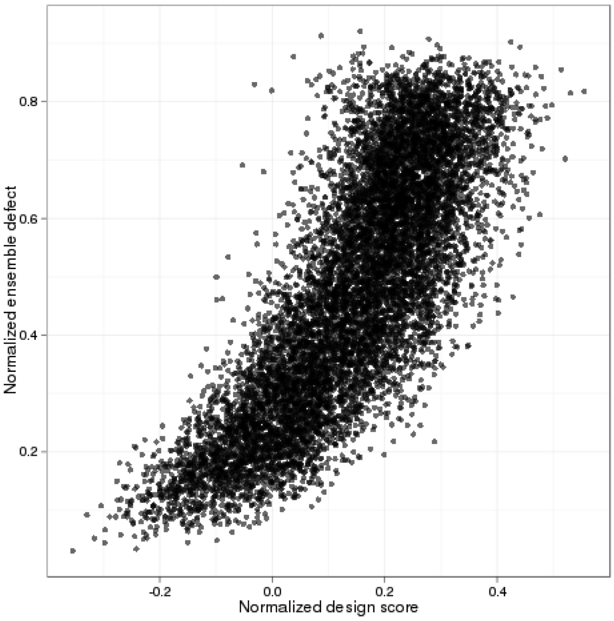

Figure 3. Comparison of normalized design scores and normalized ensemble defects for 10 000 "random compatible" sequences of sizes 100 to 3200 bases.

Quality of Designed Sequences. Given the behavior of the score function, one can then see whether sequences fold to the desired structure and when not, by how much. First, one can consider the effect of more sophisticated design methods compared to "random compatible" sequences. That is, what do the design programs bring compared to simple Watson−Crick base pairing?

Regardless of size or the type of sequence, it is clearly not of much help to simply follow base-pairing rules when designing sequences. In no case is there even any overlap between the quartile results from the design programs and the "random compatible" sequences. The results also show a property of the test set. The structures which come from folding truly random sequences are consistently more difficult to design (larger ensemble defects).

Next, one can compare the same values for the two programs. Figure 5 is the same as Figure 4, but with the results of "random compatible" sequences removed. The first thing to notice is that both programs produce sequences with very low ensemble defects. Looking at the engineered structures, the dynamics-based method produces the best typical (median) results, but with more scatter. For the structures coming from random sequences, NUPACK produces better sequences. On the basis of these results, we would not claim one program is better than the other, but it highlights a problem trying to find a fair or representative test set.

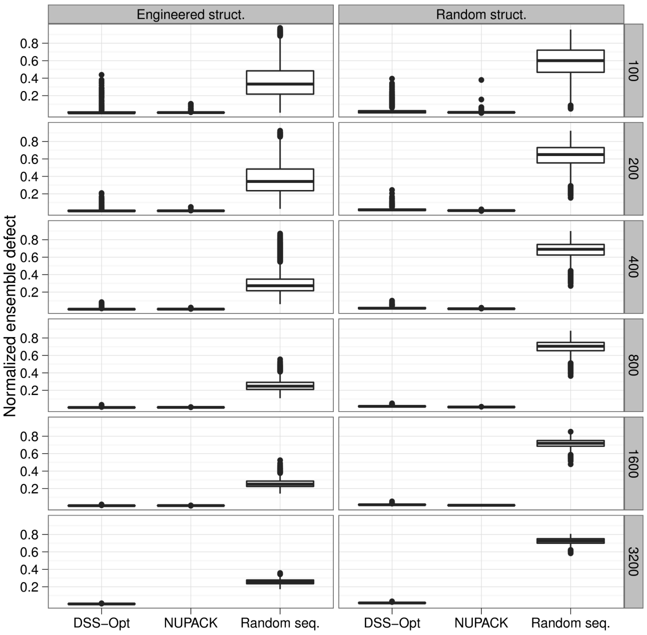

Figure 4. Box plots of normalized ensemble defect for the different programs and different parts of the test set. Plots are labeled with sequence size and program name. DSS-Opt refers to the dynamics approach. Random seq. refers to "random compatible". In each plot, the median is marked by a thicker line and the first quartiles marked. Outliers are marked individually. Results are not given for NUPACK for the larger sequences when it did not finish within the one hour time limit.

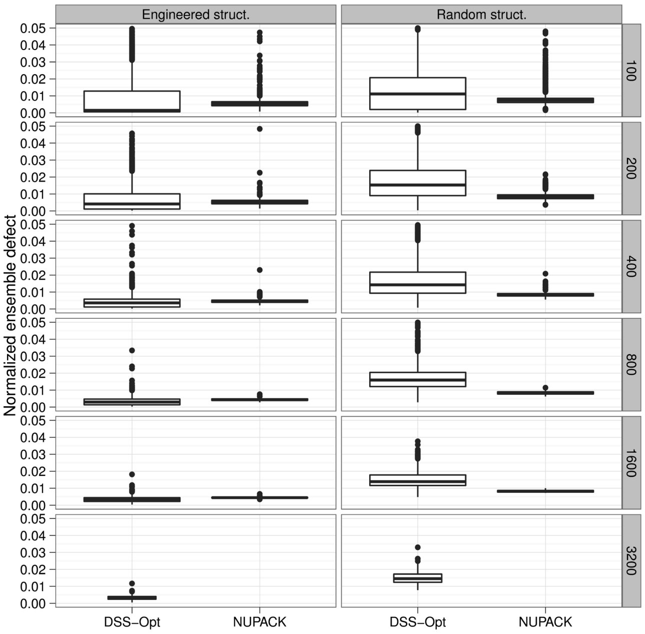

Figure 5. Box plots of normalized ensemble defect for the different design programs. Data is the same as for Figure 4, but "random compatible" sequences are omitted.

Given that both programs produce sequences which fold to the correct structure or something very close, one can ask just what is the Boltzmann probability of the lowest energy structure. Figure 6 shows results for both programs, as well as "random compatible" sequences. Clearly, "random compatible" sequences have a negligible probability of folding to the target structure. The dynamics-based method produces better sequences for the smaller, engineered structures and given the greater scatter can produce better sequences for the larger structures. Among the "random" structures, NUPACK again tends to give larger probabilities. It is interesting to note that the results show a typical feature of structure predictions calculated by methods descended from the Nussinov algorithm.[41] When a structure is large, there is always a so-called minimum free energy structure, but its probability is far from 1.0. There is a collection of slightly different, nonoptimal structures, all with nonzero probability.

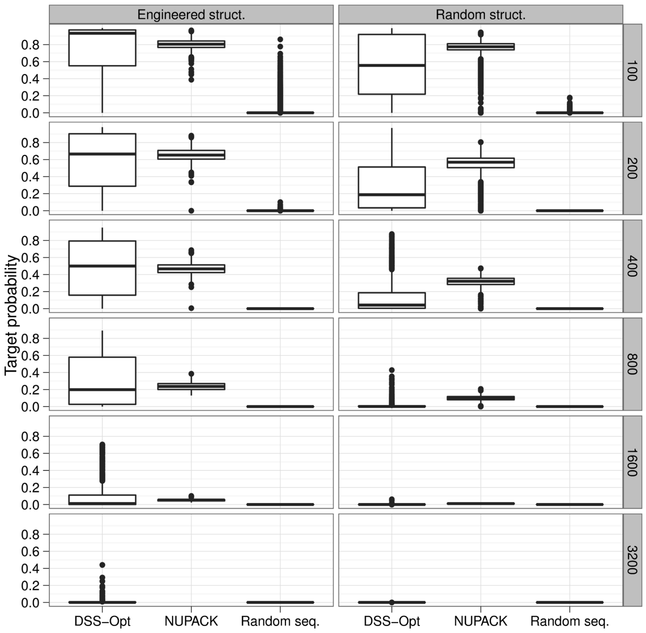

Figure 6. Target structure probability for designed sequences. Results are split by test set and labeled by method as in Figure 5. Medians, quartiles, and outliers are marked as in Figure 4.

DISCUSSION

If one views this purely as an optimization problem, it is not surprising that the method works rather well. Newtonian dynamics with temperature annealing is not an unusual optimization technique. Of course, this comes with the caveat that one cannot claim to have found an optimal solution. The interesting aspects of this problem are just how much simpler it is than a protein calculation and how one can implement the idea of negative design.

Working with a simple score function rather than an atomistic force field has some distinct advantages, albeit with a distinct cost in terms of realism. For example, moving between base types in an atomistic model would involve making or removing place or adjusting backbone angles when moving between purines and pyrimidines. In contrast, the model used here implicitly assumes that the molecule's geometry can adapt so as to improve base-pairing and base-stacking, the features which dominate the calculations. In practice, this is well-grounded. The model is based solely on experimental melting temperatures regardless of the underlying atomistic basis.[42] At the same time, the approach is too simple to claim one can really design structures. Although one would expect designed sequences to fold to the right secondary structure, there is no guarantee that two molecules with the same secondary structure will fold to the same shapes in solution.

One should also note that the nearest-neighbor model is usually referred to as a free energy approximation. It could be a bad approximation since it is a purely additive scheme, but free energies are not simple additive quantities.[43] At the same time, this is not a problem in this context. One has a scoring scheme which is parametrized against experimental values for stability. One may label this as a free energy approximation if one wants.

The real benefit of the model is the ability to implement a negative design term. That is, one numerically discourages unwanted structures. This would be extremely difficult to do with a more detailed force field. If one believes the structure probabilities we calculated using a standard method,[12] the approach is very successful. Not only is the desired structure the preferred structure for the target sequences, but it is also a highly populated structure. That is, alternative folding has been actively discouraged. At the same time, we calculated the ensemble defect[40] in the structures predicted for the designed sequences. This is of interest for two reasons. First, it emphasizes a different kind of error (unwanted base pairing), and second it was not part of the score/quasi-energy function. One could try to adjust parameters so as to produce the smallest ensemble defect measure, but nonlinear regression did not show any strong correlation between any of our adjustable parameters and ensemble defect (results not shown). This means that if one were to decide that this measure is important, the most practical approach would be an initial sampling using dynamics, followed by a brief, local optimization based on eq 12 or even scoring with a different program.[15] This naturally leads to the comparison of the dynamics-based approach and the comparison with the program NUPACK.

There has been little attempt to directly compare the programs. Both have adjustable parameters and can be made to appear better or worse, usually at the cost of some calculation time. Both programs were simply given the same amount of CPU time for each sequence. This would penalize a method if it were slower but ultimately yielded better results. The programs do implement quite different philosophies. In our method, incorrect base pairs are discouraged by a type of average interaction over unwanted base pairs. NUPACK implements a conventional discrete search and explicitly optimizes for the ensemble defect. Our approach should work well in terms of finding sequences where the target has a good estimated free energy, but it seems to work relatively well when tested in terms of the ensemble defect. This suggests the scatter in Figure 3 might be larger than one wants, but this is not a problem. There was a completely unexpected result with respect to the parts of the test set labeled "engineered" and "random". The test set was chosen because it is large and should be objective. Surprisingly, it seems that our method produces slightly better results for the "engineered" structures, whereas NUPACK is better for those structures labeled "random". Pierce et al.[15] generated the "random" set by generating random sequences and predicting their secondary structure. The "engineered" set was built with restraints to keep loop numbers and lengths similar to those used in nucleic acid nanotechnology. The "random" set then has more smaller loop regions. Apparently, the biases of the two methods differ, but it raises another issue in comparisons. Not only is it possible to adjust parameters for the different methods, the average properties of the test sets also affect the results. Another curious fact is that a convergence term forcing the system to reach a physically realizable composition, i.e., one where each base only has one base type with probability 1 and all other base types have probability 0, is not needed. The system converges to "pure" sequences by itself. This has already been seen when treating DNA sequences with a continuous representation.[22] One could compare the results using a more abstract method which checks for the presence of secondary structure elements.[44] Instead, we have assumed that a small ensemble defect means the secondary structure is mostly correct, while larger defects are too wrong to be interesting.

Looking at our score function, one would concede it has some arbitrary components. Unlike many other sequence design methods, we have an explicit heterogeneity term (eq 7). This simply means that a sequence vector at one site tends to dissuade neighboring sites from pointing in the same direction. This is hard to justify objectively. It means that when choosing between sequences, one will be preferred when it does not have stretches of identical residues. This certainly makes sequences look more biological, but this is not a strong justification. Making a sequence more heterogeneous will, however, have an effect on the energy landscape in real space. It will mean that there is less scope for nearly identical interactions and in this sense will remove some of the symmetries in the solutions. One could also note that some amount of heterogeneity will be found automatically. Stretches of similar residues lead to the possibility of "slippage," which means bases find alternative pairings with similar energies. In our formulation, the negative design term already weakly discourages this. In other design methods which directly optimize the probability of the target structure or the ensemble defect, stretches of identical residues would immediately be identified as detrimental.

With respect to dynamics in sequence space, the most relevant comparisons are with λ-dynamics and perhaps the methods coined θ-dynamics,[45] but there is a philosophical difference. In θ-dynamics, the mixture of Hamiltonians is parametrized in terms of an additional variable θ such that the sum of λ contribution is rigidly held at 1. The single θ value determines the mixing of Hamiltonians. In the sequence-design context, one could see this as forcing the particles to travel on the surface of the hypersphere where the base composition sums to exactly 1 for each site. In contrast to the problems tackled by θ-dynamics, we are not looking at comparing a small number of states but instead searching in a much larger space described by all the λ particles. In other words, the emphasis is on searching sequence space. In this case, one prefers to use soft restraints on the system so as to make barriers in the sequence space easier to surmount. This means that there is an additional parameter, k_{cmp}, which was simply set so as to be high, but not so high as to cause numerical problems with the integrator.

Finally, one must note that no method which only looks at final structures can guarantee to be successful. The kinds of functions used here do not consider issues of folding kinetics at all. It may be that a sequence is in some sense optimal for a structure, but folding to that structure could be very slow and much slower than folding to some nonoptimal but stable structure.[46]

The next step will be to synthesize RNA sequences. At the moment, one may not be able to design biologically functional molecules, but for applications such as molecular scaffolds, DSS-Opt may be a useful procedure.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: matthies@zbh.uni-hamburg.de; torda@zbh.uni-hamburg.de.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Professor van Gunsteren for turning 65, developing many entertaining algorithms, and provoking us to use them.

■ ADDITIONAL NOTE

DSS-Opt is freely available under the terms of the GNU General Public License (GPL) at http://github.com/marcom/dss-opt/.

■ REFERENCES

(1) Mathews, D. H.; Turner, D. H. Prediction of RNA secondary structure by free energy minimization. *Curr. Opin. Struct. Biol.* **2006**, *16*, 270−278.

(2) Floudas, C. A.; Fung, H. K.; McAllister, S. R.; Mönnigmann, M.; Rajgaria, R. Advances in protein structure prediction and de novo protein design: A review. *Chem. Eng. Sci.* **2006**, *61*, 966−988.

(3) Samish, I.; MacDermaid, C. M.; Perez-Aguilar, J. M.; Saven, J. G. Theoretical and Computational Protein Design. *Annu. Rev. Phys. Chem.* **2011**, *62*, 129−149.

(4) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953**, *21*, 1087.

(5) Hellinga, H. W.; Richards, F. M. Optimal sequence selection in proteins of known structure by simulated evolution. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 5803−5807.

(6) Dantas, G.; Kuhlman, B.; Callender, D.; Wong, M.; Baker, D. A Large Scale Test of Computational Protein Design: Folding and Stability of Nine Completely Redesigned Globular Proteins. *J. Mol. Biol.* **2003**, *332*, 449−460.

(7) Kuhlman, B.; Dantas, G.; Ireton, G. C.; Varani, G.; Stoddard, B. L.; Baker, D. Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Science* **2003**, *302*, 1364−1368.

(8) Kirkpatrick, S.; Gelatt, C. D., Jr.; Vecchi, M. P. Optimization by Simulated Annealing. *Science* **1983**, *220*, 671−680.

(9) Cootes, A. P.; Curmi, P. M. G.; Torda, A. E. Biased Monte Carlo optimization of protein sequences. *J. Chem. Phys.* **2000**, *113*, 2489−2496.

(10) Zou, J.; Saven, J. G. Using self-consistent fields to bias Monte Carlo methods with applications to designing and sampling protein sequences. *J. Chem. Phys.* **2003**, *118*, 3843−3854.

(11) Yang, X.; Saven, J. G. Computational methods for protein design and protein sequence variability: biased Monte Carlo and replica exchange. *Chem. Phys. Lett.* **2005**, *401*, 205−210.

(12) Hofacker, I. L.; Fontana, W.; Stadler, P. F.; Bonhoeffer, L. S.; Tacker, M.; Schuster, P. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.* **1994**, *125*, 167−188.

(13) Andronescu, M.; Fejes, A. P.; Hutter, F.; Hoos, H. H.; Condon, A. A New Algorithm for RNA Secondary Structure Design. *J. Mol. Biol.* **2003**, *336*, 607−624.

(14) Busch, A.; Backofen, R. INFO-RNA — a fast approach to inverse RNA folding. *Bioinformatics* **2006**, *22*, 1823−1831.

(15) Zadeh, J. N.; Wolfe, B. R.; Pierce, N. A. Nucleic acid sequence design via efficient ensemble defect optimization. *J. Comput. Chem.* **2011**, *32*, 439−452.

(16) Desmet, J.; de Maeyer, M.; Hazes, B.; Lasters, I. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* **1992**, *356*, 539−542.

(17) Desmet, J.; Spriet, J.; Lasters, I. Fast and accurate side-chain topology and energy refinement (FASTER) as a new method for protein structure optimization. *Proteins* **2002**, *48*, 31−43.

(18) Allen, B. D.; Mayo, S. L. Dramatic performance enhancements for the FASTER optimization algorithm. *J. Comput. Chem.* **2006**, *27*, 1071−1075.

(19) Dahiyat, B. I.; Mayo, S. L. De Novo Protein Design: Fully Automated Sequence Selection. *Science* **1997**, *278*, 82−87.

(20) Koehl, P.; Levitt, M. De novo protein design. I. In search of stability and specificity. *J. Mol. Biol.* **1999**, *293*, 1161−1181.

(21) Voigt, C. A.; Gordon, D. B.; Mayo, S. L. Trading accuracy for speed: a quantitative comparison of search algorithms in protein sequence design. *J. Mol. Biol.* **2000**, *299*, 789−803.

(22) Lafontaine, I.; Lavery, R. Optimization of Nucleic Acid Sequences. *Biophys. J.* **2000**, *79*, 680−685.

(23) Hu, X.; Hu, H.; Beratan, D. N.; Yang, W. A gradient-directed Monte Carlo approach for protein design. *J. Comput. Chem.* **2010**, *31*, 2164−2168.

(24) Tidor, B. Simulated annealing on free energy surfaces by a combined molecular dynamics and Monte Carlo approach. *J. Phys. Chem.* **1993**, *97*, 1069−1073.

(25) Kong, X.; Brooks, C. L. λ-dynamics: A new approach to free energy calculations. *J. Chem. Phys.* **1996**, *105*, 2414.

(26) Godzik, A. In search of the ideal protein sequence. *Protein Eng.* **1995**, *8*, 409−416.

(27) Torda, A. E. Protein Sequence Optimization—Theory, Practice, and Fundamental Impossibility. *Soft Mater.* **2004**, *2*, 1−10.

(28) Jones, D. T. De novo protein design using pairwise potentials and a genetic algorithm. *Protein Sci.* **1994**, *3*, 567−574.

(29) Shakhnovich, E. I.; Gutin, A. M. A new approach to the design of stable proteins. *Protein Eng.* **1993**, *6*, 793−800.

(30) Hom, G. K.; Mayo, S. L. A search algorithm for fixed-composition protein design. *J. Comput. Chem.* **2006**, *27*, 375−378.

(31) Seno, F.; Vendruscolo, M.; Maritan, A.; Banavar, J. R. Optimal Protein Design Procedure. *Phys. Rev. Lett.* **1996**, *77*, 1901−1904.

(32) Bhattacherjee, A.; Biswas, P. Statistical Theory of Protein Sequence Design by Random Mutation. *J. Phys. Chem. B* **2009**, *113*, 5520−5527.

(33) Chiu, T. L.; Goldstein, R. A. Optimizing potentials for the inverse protein folding problem. *Protein Eng.* **1998**, *11*, 749−752.

(34) Knight, J. L.; Brooks, C. L., III. λ-Dynamics Free Energy Simulation Methods. *J. Comput. Chem.* **2009**, *30*, 1692−1700.

(35) Darty, K.; Denise, A.; Ponty, Y. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics* **2009**, *25*, 1974−1975.

(36) Hockney, R. W.; Eastwood, J. W. *Computer Simulation Using Particles*; McGraw-Hill: New York, 1988; p 94.

(37) Harvey, S. C.; Tan, R. K.-Z.; Cheatham, T. E., III. The Flying Ice Cube: Velocity Rescaling in Molecular Dynamics Leads to Violation of Energy Equipartition. *J. Comput. Chem.* **1998**, *19*, 726−740.

(38) José, J. V.; Saletan, E. J. *Classical Dynamics: A Contemporary Approach*; Cambridge University Press: Cambridge, U. K., 1998; p 251.

(39) McCaskill, J. S. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **1990**, *29*, 1105−1119.

(40) Dirks, R. M.; Lin, M.; Winfree, E.; Pierce, N. A. Paradigms for computational nucleic acid design. *Nucleic Acids Res.* **2004**, *32*, 1392−1403.

(41) Nussinov, R.; Pieczenik, G.; Griggs, J. R.; Kleitman, D. J. Algorithms for Loop Matchings. *SIAM J. Appl. Math.* **1978**, *35*, 68−82.

(42) Mathews, D. H.; Sabina, J.; Zuker, M.; Turner, D. H. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* **1999**, *288*, 911−940.

(43) Dill, K. A. Additivity Principles in Biochemistry. *J. Biol. Chem.* **1997**, *272*, 701−704.

(44) Giegerich, R.; Voß, B.; Rehmsmeier, M. Abstract shapes of RNA. *Nucleic Acids Res.* **2004**, *32*, 4843−4851.

(45) Knight, J. L.; Brooks, C. L., III. Applying efficient implicit nongeometric constraints in alchemical free energy simulations. *J. Comput. Chem.* **2011**, *32*, 3423−3432.

(46) Flamm, C.; Hofacker, I. L. Beyond energy minimization: approaches to the kinetic folding of RNA. *Monatsh. Chem.* **2008**, *139*, 447−457.