Software News and Update
NUPACK: Analysis and Design of Nucleic Acid Systems

JOSEPH N. ZADEH,[1†] CONRAD D. STEENBERG,[1†] JUSTIN S. BOIS,[1†] BRIAN R. WOLFE,[1†] MARSHALL B. PIERCE,[1]
ASIF R. KHAN,[1] ROBERT M. DIRKS,[1] NILES A. PIERCE[1,2]

[1]*Department of Bioengineering, California Institute of Technology, Pasadena, California, 91125*
[2]*Department of Applied and Computational Mathematics, California Institute of Technology, Pasadena, California, 91125*

Received 4 April 2009; Accepted 10 May 2010
DOI 10.1002/jcc.21596
Published online in Wiley InterScience (www.interscience.wiley.com).

Abstract: The Nucleic Acid Package (NUPACK) is a growing software suite for the analysis and design of nucleic acid systems. The NUPACK web server (`http://www.nupack.org`) currently enables:

- Analysis: thermodynamic analysis of dilute solutions of interacting nucleic acid strands.
- Design: sequence design for complexes of nucleic acid strands intended to adopt a target secondary structure at equilibrium.
- Utilities: evaluation, display, and annotation of equilibrium properties of a complex of nucleic acid strands.

NUPACK algorithms are formulated in terms of nucleic acid secondary structure. In most cases, pseudoknots are excluded from the structural ensemble.

 J Comput Chem 000: 000–000, 2010

Key words: RNA; DNA; complexes; concentrations; sequence design

Introduction

The analysis and design of RNA and DNA base-pairing arise both in basic biological research and in engineering novel chemical and biological systems. NUPACK is focused on the analysis and design of nucleic acid secondary structure for systems involving one or more species of interacting strands. Notable features include:

- Calculation of the partition function and minimum free energy (MFE) secondary structure for unpseudoknotted complexes of arbitrary numbers of interacting RNA or DNA strands* including rigorous treatment of distinguishability issues that arise in the multi-stranded setting.[1]
- Calculation of the equilibrium concentrations for arbitrary species of complexes in a dilute solution (e.g., for a test tube of interacting RNA or DNA strand species).[1]
- Use of partition function and concentration information to calculate equilibrium base-pairing observables for dilute solutions of interacting strand species.[1]
- Sequence design for one or more strands intended to adopt an unpseudoknotted target secondary structure at equilibrium.[2]

The NUPACK web server also offers utilities for the customization of figures for talks and papers, providing:

- Publication-quality vector graphics that can be downloaded and edited in standard vector graphics programs.
- Automatic layout and rendering of secondary structures depicted with or without ideal helical geometry.

*The web server currently limits the maximum complex size to ten strands and displays a single MFE structure. Larger complexes and degenerate MFE structures can be analyzed by downloading and compiling the NUPACK source code.

†These authors contributed equally to this work.

Correspondence to: N.A. Pierce; Email: niles@caltech.edu

Contract/grant sponsor: National Science Foundation (The Molecular Programming Project); contract/grant number: CCF-0832824 , CHE-0533064, DMS-0506468

Contract/grant sponsor: The National Institutes of Health; contract/grant number: P50 HG004071

Contract/grant sponsor: The Ralph M. Parsons Foundation

Contract/grant sponsor: Beckman Institute at Caltech

- Dynamic graphical editing of secondary structure layout within the web interface.

This note summarizes the features of the Analysis, Design, and Utilities pages of the NUPACK web server.

Secondary Structure Model

The *secondary structure* of multiple interacting strands is defined by a list of base pairs.[1] A *polymer graph* for a secondary structure can be constructed by *ordering* the strands around a circle, drawing the backbones in succession from 5′ to 3′ around the circumference with a *nick* between each strand and drawing straight lines connecting paired bases. A secondary structure is *pseudoknotted* if every strand ordering corresponds to a polymer graph with crossing lines. A secondary structure is *connected* if no subset of the strands is free of the others. Algorithms are formulated in terms of *ordered complexes*, each corresponding to the structural ensemble of all connected polymer graphs with no crossing lines for a particular ordering of a set of strands. The free energy of an unpseudoknotted secondary structure is calculated using nearest-neighbor empirical parameters for RNA in 1M Na^+[3,4] or for DNA in user-specified Na^+ and Mg^{++} concentrations[5–7]; additional parameters are employed for the analysis of pseudoknots (single RNA strands only).[8,9]

Analysis

The Analysis page allows users to analyze the thermodynamic properties of a dilute solution of interacting nucleic acid strands in the absence of pseudoknots.

Input

The Analysis Input page allows the user to specify the components and conditions of the solution of interest: RNA or DNA, temperature (or range of temperatures for melts), number of strand species, maximum complex size (all ordered complexes with up to this number of strands will be included in the analysis), strand sequences, strand concentrations (for calculations with maximum complex size greater than one). Under the expandable Advanced Options panel, users may select among available energy models, specify salt concentrations, allow a class of pseudoknots (single RNA strands only), and specify additional ordered complexes to include in the calculation (larger than the specified maximum complex size). The estimated computation time is displayed as the user provides input. If the estimate exceeds a threshold, an email address is required to permit notification of job completion; jobs estimated to exceed a larger threshold are not accepted.

Computation

The partition function, equilibrium base-pairing probabilities, and MFE structure are calculated for each ordered complex using dynamic programs.[1,8,9] For calculations in which the maximum complex size is greater than one, the calculated partition functions and user-provided strand concentrations are used to calculate the equilibrium concentration of each ordered complex by solving a convex optimization problem.[1] If the user wishes to change the strand concentrations after examining the results, it is not necessary to recompute the partition functions, and the equilibrium properties of the dilute solution can be rapidly recomputed from within the Results page.

Results

The Results page summarizes the equilibrium properties of the dilute solution:

- Melt profile plot: Depicts the fraction of unpaired bases at equilibrium as a function of temperature.
- Ensemble pair fractions plot: Depicts equilibrium base-pairing information for the dilute solution, taking into account the equilibrium concentration and base-pairing properties of each ordered complex. Each entry in the plot provides information about a particular species of base pair (e.g., the base pair in which base i of strand species A (row) pairs to base j of strand species B (column); the color and area of the corresponding dot scale with the fraction of strands of species A that form this pair at equilibrium). In general, the matrix is not symmetric. Each dot in the column at right represents the fraction of strands of a given species with the corresponding base unpaired at equilibrium.
- Equilibrium concentration histogram: Depicts the equilibrium concentrations of the ordered complexes.

Clicking on any bar in the histogram displays equilibrium information about the corresponding ordered complex:

- MFE structure plot: Depicts the MFE secondary structure for the ordered complex. In the default view, each base is shaded with the probability that it adopts the depicted paired or unpaired state at equilibrium, allowing the user to assess the utility of different portions of the MFE structure in summarizing the structural features of the ordered complex ensemble.
- Pair probabilities plot: Depicts equilibrium base-pairing probabilities for the ordered complex, treating all strands as distinct. By definition, these data are independent of concentration and of all other ordered complexes in solution. The area and color of each dot scale with the equilibrium probability of the corresponding base pair. With this convention, the plot is symmetric, with the upper and lower triangles separated by a diagonal line. The area and color of each dot in the column at right scale with the equilibrium probability that the corresponding base is unpaired. Optional black circles depict each base pair or unpaired base in the MFE structure.

The MFE structure can be exported to the Design page (e.g., to redesign the sequence), and the MFE structure and strand sequences can be exported to the Utilities page (e.g., to annotate or edit publication-quality graphics). The MFE and pair probability images may be downloaded in SVG format for editing in vector graphics programs. Alternatively, all data and plots can be downloaded as a single compressed file.

Design

The Design page allows users to design sequences for one or more strands intended to adopt an unpseudoknotted target secondary structure at equilibrium.

Input

The Design Input page allows the user to specify design requirements: RNA or DNA, temperature, number of independent sequence designs, target secondary structure in *dot-parens-plus notation* (each unpaired base is represented by a dot, each base pair by matching parentheses, and each nick between strands by a plus). Target secondary structures that are multi-stranded must be connected. Under the expandable Advanced Options panel, users may select among available energy models, specify salt concentrations, specify sequence constraints, and define pattern prevention requirements.

Computation

The design algorithm performs efficient ensemble defect optimization to reduce the *ensemble defect*,[2] the average number of incorrectly paired nucleotides at equilibrium calculated over the structural ensemble of the ordered complex.[10] For a target secondary structure with N nucleotides, the algorithm seeks to achieve an ensemble defect below $N/100$.

Results

The Results page summarizes the properties of the designed sequences:

- Designability summary: Depicts each base in the target secondary structure shaded by the probability that it adopts the depicted paired or unpaired state at equilibrium, averaged across the independent sequence designs. This plot can expose conceptual design flaws in the target structure: if a particular base pair has a low probability of forming over several independent sequence designs, adjustments to the target structure may be warranted.
- Sequence designs table: Displays the ensemble defect, normalized ensemble defect, GC content, and sequences for each design.

Clicking on a sequence design displays equilibrium information about the ordered complex to which the target secondary structure belongs:

- Target structure plot: Depicts the target secondary structure with each base shaded by the probability that it adopts the depicted state at equilibrium.
- Pair probabilities plot: Depicts the equilibrium base-pairing probabilities for the ordered complex. Optional black circles depict each base pair or unpaired base in the target secondary structure.

Any set of designed sequences can be exported to the Analysis page (e.g., to check for the formation of unintended ordered complexes in the context of a dilute solution). The target structure and any set of designed sequences can be exported to the Utilities page (e.g., to customize publication-quality graphics).

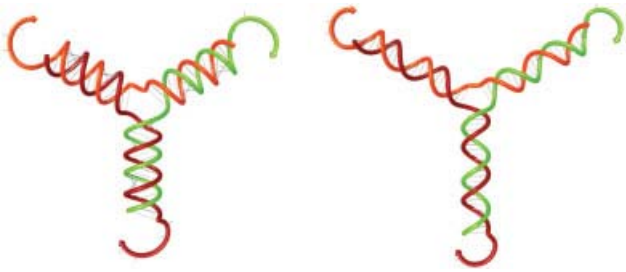

Figure 1. The Utilities page enables depiction of secondary structures with ideal helical geometry for stacked base pairs, as for this complex of three RNA strands with A-form helices **(left)** or three DNA strands with B-form helices **(right)**.

Utilities

The Utilities page allows users to evaluate, display, and annotate the equilibrium properties of an ordered complex. The page accepts as input either sequence information, structure information, or both, performing diverse functions based on the information provided, including:

- Evaluation and display of equilibrium information for a specified secondary structure in the context of the ordered complex to which it belongs (analogous to the treatment of the MFE structure in the Analysis page, and the target structure in the Design page).
- Automatic layout and rendering of secondary structures specified in dot-parens-plus notation with (e.g., see Fig. 1) or without ideal helical geometry. In either case, the structure layout can be edited dynamically within the web application.

Sequences can be exported to the Analysis page for further examination in the context of a dilute solution. Alternatively, structures can be exported to the Design page, carrying any specified sequence information as design constraints.

Implementation

The NUPACK web application is programmed within the Ruby on Rails framework, employing AJAX and the Dojo Toolkit to implement dynamic features and interactive graphics. Plots and graphics are generated using NumPy and matplotlib. The site is supported on current versions of the Safari, Chrome, and Firefox browsers. The NUPACK library of analysis and design algorithms is written in the C programming language. Dynamic programs are parallelized using MPI.[11] The NUPACK web server is for noncommercial research purposes.

Acknowledgments

The authors W. Yardley, N. Near-Ansari, and C. Schmutzer for technical support in deploying the NUPACK server, M. O'Connell for programming assistance, and V.A. Beck, J.D. Bishop, H.M.T. Choi, E. Franco, L.M. Hochrein, P.W.K. Rothemund, M. Schwarzkopf, J.B. Sternberg, S. Venkataraman, J.R. Vieregg, E. Winfree, P. Yin, and D.Y. Zhang for comments and testing.

References

1. Dirks, R. M.; Bois, J. S.; Schaeffer, J. M.; Winfree, E.; Pierce, N. A., SIAM Rev 2007, 49, 65.
2. Zadeh, J. N.; Wolfe, B. R.; Pierce, N. A. J Comput Chem, in press.
3. Serra, M. J.; Turner, D. H. Methods Enzymol 1995, 259, 242.
4. Mathews, D. H.; Sabina, J.; Zuker, M.; Turner, D. H. J Mol Biol 1999, 288, 911.
5. SantaLucia, Jr. J. Proc Natl Acad Sci USA 1998, 95, 1460.
6. SantaLucia, J.; Hicks, D. Annu Rev Bioph Biom Struct 2004, 33, 415.
7. Koehler, R. T.; Peyret, N., Bioinformatics 2005, 21, 3333.
8. Dirks, R. M.; Pierce, N. A. J Comput Chem 2003, 24, 1664.
9. Dirks, R. M.; Pierce, N. A. J Comput Chem 2004, 25, 1295.
10. Dirks, R. M.; Lin, M.; Winfree, E.; Pierce, N. A. Nucleic Acids Res 2004, 32, 1392.
11. Fekete, M.; Hofacker, I. L.; Stadler, P. F. J Comput Biol 2000, 7, 171.