Can't (under)stand the heat?

As genomic assays become more complex, hybridization software becomes hotter.

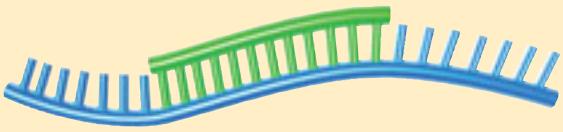
BY ROBERT ROYCE

The demand for assays to detect and quantify specific DNA or RNA sequences is growing rapidly. Whether screening blood for pathogens, characterizing gene transcription patterns, or studying the relationship between genotype and phenotype, scientists are continually looking for ways to efficiently scale their investigations and refine their measurements. Generally, these genomic assays require carefully designed oligonucleotide primers and probes that are mixed with genomic samples to detect DNA or RNA targets.

Ideally, researchers would be able to routinely design complex assays, such as multiplex PCR reactions that work the first time, and to devise specific and sensitive microarray experiments with minimal trial and error. Until recently, this simply wasn't possible. Designing complex assays requires more than the simple melting-temperature calculators found in typical oligonucleotide design software (Figure 1).

On the basis of the wide range of free and inexpensive oligonucleotide design software and Web-based design services, one might think that the task is easy. After all, nucleic acids are some of the easiest biological molecules to control in vitro. The strong affinity between G–C and A–T (A–U for RNA), and the weaker affinity for other combinations of purines (G, A) and pyrimidines (C, T, U), simplifies the problem considerably.

However, designing specific and sensitive assays of any complexity is still difficult. Oligonucleotides tend to partially bind to other DNA and RNA molecules, leading to false positives, and target molecules fold upon themselves, which inhibits binding, leading to false negatives. In addition, many assays have special requirements such as allele specificity or probes that must be fixed in particular locations. Undesired primer–primer or primer–probe interactions further complicate the design process. While each of these details affects the outcome, programs that simply focus on calculating the melting temperature fail to account for these artifacts.



$$T_M = 2(A + T) + 4(G + C) - 16.6 \log [Na^+]$$

Figure 1. That was then. Using a simple equation, researchers get a very rough estimate of the melting temperature of a nucleic acid duplex.

In the end, what a researcher really wants to know is not the temperature at which an oligonucleotide "melts" into a random coil but rather the binding efficiency between a primer/probe and the target and what artifacts are formed during the reaction. So while free software and services may suffice for simple experiments, any significant effort must take into account the many factors contributing to hybridization, such as folding, cross-hybridization, and the presence of magnesium or sodium. This is why in silico modeling techniques provide superior results to those obtained from the simplistic calculations used in the most popular probe-and-primer design software.

It is only natural that in silico modeling would find an application in the biology laboratory. Manufacturers of everything from nanotubes to airplanes have adapted in silico modeling techniques to automate the design of complex parts and systems. Applying these techniques to genomics promises to enable "industrial" genomics, with gains in flexibility, scale, and precision.

A comprehensive in silico model goes beyond calculating melting temperature, allowing scientists to simulate their experiments ahead of time. Rather than ordering thousands of primers and experimentally testing every possible design in vitro, as is often done, researchers can simulate the same process using an in silico model, quickly reducing the problem space. For most applications, an accurate model will select an excellent design that works, but there is always the option of comparing the best prospects using traditional experimental methods.

Modeling keys

Any computer model must address the trade-offs between the desire for perfection and the need for expediency. Striking the right balance is essential to practicality. The key is making the right a priori assumptions about what is relevant and what is not. More detail may improve precision, but at the cost of both model-building and computational time. For example, detailed molecular dynamics simulations of biomolecules yield great insight into the structural basis for molecular interactions and behavior, yet the simulations are computationally expensive. Simply simulating a nucleic acid with a few water molecules and cations for a few nanoseconds can take weeks or months on a supercomputer.

At the other end of the spectrum are simple mathematical models based on general parameters, such as counting the number of Gs and Cs in a primer to estimate its melting temperature when bound to a target. This type of calculation is fast but inaccurate (typical error >10 °C). When amplifying one uncomplicated target sequence, this simplistic model often works fine when there is the luxury of experimentally optimizing the annealing temperature, magnesium concentration, and primer concentrations. For multiplex PCR, however, all amplifications occur under the same con-

ditions, and the large number of variables makes it very difficult to optimize experimentally. Differential amplification of the multiple targets and cross-hybridization artifacts becomes inevitable. In silico design allows for these artifacts to be anticipated, resulting in much higher success rates.

The key to striking the right balance is to truly understand the physical nature of what is being modeled. At the heart of most oligonucleotide design tools is a basic abstraction first developed for RNA by chemistry professor Ignacio Tinoco, Jr., at the University of California, Berkeley, in the 1970s and later refined and expanded by chemistry professor Douglas Turner at the University of Rochester. This nearest-neighbor model abstracts the covalent structure in a nucleic acid duplex and assigns an experimentally determined energy contribution to each possible quartet of two adjoining base pairs. Thus, a string of "nearest neighbors" is combined with an initiation parameter to predict the thermodynamic energy of the entire duplex. This is a very useful abstraction that avoids computing 3D structure, solvation, and electrostatics.

Figure 2. And this is now. Using software that recognizes multiple nucleic acid complexes, researchers can determine the melting temperature, T_m, very accurately. (Image courtesy of DNA Software.)

In the late 1990s, John SantaLucia, a chemistry professor at Wayne State University and chief scientist at DNA Software, refined the nearest-neighbor model for DNA and systematically determined a database of thermodynamic parameters that account for a wide variety of phenomena such as mismatches, dangling ends, and loops (*1, 2*).

An important advance developed by SantaLucia, which is proprietary to DNA Software, is the use of a multistate thermodynamic model. Most primer-and-probe design software calculates simple two-state thermodynamics that assumes species go directly from random coils to hybridized duplex. This simplified model ignores important interactions that occur within and between molecules. A multistate thermodynamic model takes into account the energy needed to unfold sequences and the competition between the desired "perfect match" and the possible mis-hybridization that might occur. This competition can now be simulated in silico and the concentrations of all the competing species numerically solved for, including the desired hybridization and the undesired artifact signals. The result is a more accurate calculation that serves as a fundamental building block of detailed models.

The ultimate challenge of in silico modeling is taking the collection of functions that characterize the system and building them into a framework that supports flexible investigation and analysis. Software freely available for calculating thermodynamics contains a number of useful functions, but these cannot be used alone to design a complex assay.

On the other hand, Visual OMP, the system developed by SantaLucia and colleagues at DNA Software, builds the fundamental scientific principles into an in silico laboratory for simulating and designing DNA- and RNA-based experiments. Building on the foundation of SantaLucia's more accurate thermodynamic model, DNA Software programmers have built a complete workbench for designing experiments (Figure 2).

To design an experiment, whether for microarrays, multiplex PCR, or TaqMan or Scorpion assays, scientists import or enter the required sequences, including all targets and any fixed probes or primers previously identified. They then specify buffer, solvents, temperature, and concentrations of species.

An integrated folding engine allows scientists to detect strong hairpins and to select the best binding sites on long targets, which saves considerable time over accessing another algorithm, MFOLD, over the Internet. It also helps researchers avoid any potential licensing or security problems associated with commercial use. BLAST functionality is built in so that researchers can automatically design primers that avoid errant hybridization against specified genomes.

The competition between the desired "perfect match" and the possible mis-hybridization that might occur can now be simulated in silico.

Once the researcher determines an oligonucleotide design, he or she can use the results to simulate an entire experiment, viewing the process through a variety of graphical interfaces to depict cross-hybridization or to analyze temperature sensitivity by graphing the predicted concentration of every possible species.

A Developer edition of OMP is available for bioinformaticists who want to conduct large-scale simulations—for example, simulating millions of probe candidates. To support this function, researchers can use the system to take experimental designs as inputs (using a scriptable keyword interface) and return thermodynamic predictions as text files that can be integrated into a digital workflow without human intervention.

Design in practice

Recently, DNA Software researchers used OMP to design a multiplex PCR reaction with six targets for EraGen Biosciences, a biotechnology company focused on new technology for molecular diagnostics and drug discovery.

"This was an assay I attempted to design with my current tools, but the results required significant redesign. I was impressed that OMP's in silico model was able to immediately account for, and design around, a variety of problems that would

normally have taken several days or more of experimentation to resolve," says Scott Johnson, senior scientist at EraGen. The company's assays contained as many as 16 target regions in a single well, with multiple interrogations per amplicon. The assays covered a broad range of applications, including clinical diagnostics, bacterial and viral detection, and bacterial subtyping. "Since EraGen's MultiCode technology requires numerous DNA oligonucleotides, the ability to accurately predict and avoid deleterious interactions is clearly a highlight of this method," Johnson adds.

Similarly, researchers at Biocept, a developer of low- to medium-density microarray technology, used the system to model secondary target structures and facilitate the design of microarrays and multiplex PCR primers.

DNA Software's scientific and software teams continue to develop in silico models and will shortly release a new version with the added ability to design molecular beacons and account for the effects of modified bases, fluors, pH, and PCR additives, such as DMSO and formamide. The new version will also enable researchers to design degenerate primers and probes.

In silico oligonucleotide modeling systems are emerging as powerful adjuncts to traditional experimental techniques.

In silico oligonucleotide modeling systems are emerging as powerful adjuncts to traditional experimental techniques. These methods will become more useful as computing power increases and the price of computing capabilities decreases. As a result, next-generation modeling platforms hold great promise for biological researchers.

References

(1) SantaLucia, J., Jr. *Proc. Natl. Acad. Sci. U.S.A.* **1998,** *95,* 1460–1465.
(2) Peyret, N. P.; et al. *Biochemistry* **1999,** *38,* 3468–3477.

Robert Royce is a managing partner at RFD Insight and former vice president for business development at DNA Software. ■