— Keynote Address —

Predicting Nucleic Acid Hybridization and Melting Profiles

Michael Zuker
zukerm@rpi.edu
Rensselaer Polytechnic Institute, Troy, NY, USA

Many applications in modern biotechnology require a rapid and sensitive prediction of hybridization or partial hybridization between an oligonucleotide and potential targets in a genomic DNA or mRNA database. In addition, the accurate prediction of melting profiles between an oligonucleotide and a target nucleic acid is also of great value. DNA and RNA chip technologies [16, 17], PCR primer design, sequencing by hybridization and gene diagnostic methods, including SNP detection, are all technologies for which these predictions are very important. DNA chips alone have numerous applications. They are useful to monitor whole genome gene expression [15]. They are well adapted to the detection of single nucleotide polymorphisms (SNPs) [8], to identifying organisms from their sequences [2] or the characterization of splicing variants [5]. They can be used for DNA sequencing [3] or to search for protein target sites on DNA [7].

Computational methods for hybridization and melting prediction tend to make use of existing tools. Thus the very well-known BLAST [1] program is used for database searching to determine oligonucleotide specificity. BLAST or MegaBLAST are inappropriate methods since they were designed to search for similarity based on an evolutionary model rather than hybridization based on equilibrium thermodynamics. Similarly, melting temperatures for folded, single stranded oligonucleotides, or hybridized pairs of single stranded nucleic acids are usually determined using simple two state models that assume a single, presumably minimum energy folding, or hybridization *versus* a totally unfolded or non-hybridized state. In the case of hybridization, the target is assumed to be the reverse complement, with possibly a few mismatches allowed, so that no computation is needed to determine the hybridized state. In some cases, melting temperatures are estimated by using *ad hoc* methods based on GC content.

We have undertaken a research program to provide more sophisticated and therfore, we hope, more accurate algorithms and software to address the above issues. In the case of DNA or RNA melting, the work is also of theoretical interest, since we are attempting to determine entire melting profiles, and not just melting temperatures. In addition, we are attempting to determine thermodynamic values, such as the free energy, enthalpy and entropy changes during melting. These can be measured using calorimetric methods.

We have already created a prototype version of a program that we have named "FASTH" (FAST Hybridization). The name was deliberately chosen because of similarities to the existing FASTA software [9, 11, 12, 13]. A nucleic acid database is "chopped up", or hashed into a table of k-mers (words of size k), where k is chosen in advance and whose value depends on the type of search being performed. A two letter {R,Y} alphabet is used (puRine, pYrimidine) so that by always allowing R·Y base pairs, we pick up G·T/U wobble pairs along with Watson-Crick base pairs in perfect word matches. Word matches with undesired C·A pairs are discarded. Diagonals with N_c word matches or greater, where N_c is variable parameter, are marked for further analysis. For the selected diagonals,

neighboring diagonals are searched for possible word matches. This allows us to find hybridizations containing mismatches together with small bulges. An initial score based on summing nearest neighbor free energy parameters for base pair stacking is determined, and this score is used to rank the hits. A histogram of these scores is created, and the top hits are presented in an output list. These top hits are further analyzed by computing a minimum free energy hybridization for each one. These energy scores are used to reorder the list.

For detailed hybridization and melting predictions, we compute partition functions for single stranded oligonucleotides and their targets, so that all possible states are considered. The algorithm is our own version of the McCaskill algorithm [10] and its implementation in the RNAfold program [6]. Partition functions are also computed for all hybridized states, employing an algorithm that is similar in spirit to one described by Finkelstein and Roytberg [4]. In addition, we consider the oligonucleotide and its target together in a single calculation for each temperature. The possible formation of homodimers is considered along with the oligonucleotide to target interactions. Equilibrium concentrations of all single and double stranded states are computed. This allows us to compute ensemble free energies and base pair probabilities. Numerical differentiation of a free energy *versus* temperature curve yields a computed heat capacity as a function of temperature. Combining computed probabilities of pairing with measured extinction coefficients [14], gives us simulated UV absorbance curves at 260nm.

Some theoretical details will be presented along with sample calculations. These calculations will include cases where observed melting curves are available. Ideas for further research will be discussed, including the consideration of base stacking in unfolded, single stranded nucleic acids.

References

[1] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J., Basic local alignment search tool, *J. Mol. Biol.*, 215:403–410, 1990.

[2] Cho, J.C. and Tiedje, J.M., Bacterial species determination from DNA-DNA hybridization by using genome fragments and DNA microarrays, *Appl. Environ. Microbiol.*, 67(8):3677–3682, 2001.

[3] Drmanac, S., Kita, D., Labat, I., Hauser, B., Schmidt, C., Burczak, J.D., and Drmanac, R., Accurate sequencing by hybridization for DNA diagnostics and individual genomics, *Nat Biotechnol.*, 1:54–58, 1998.

[4] Finkelstein, A.V. and Roytberg, M.A., Computation of biopolymers: a general approach to different problems, *BioSystems*, 30:1–19, 1993.

[5] Ho, L., Guo, Y., Spielman, L., Petrescu, O., Haroutunian, V., Purohit, D., Czernik, A., Yemul, S., Aisen, P.S., Mohs, R., and Pasinetti, G.M., Altered expression of a-type but not b-type synapsin isoform in the brain of patients at high risk for Alzheimer's disease assessed by DNA microarray technique, *Neurosci Lett.*, 3:191–194, 2001.

[6] Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhöffer, S., Tacker, M., and Schuster, P., Fast folding and comparison of RNA secondary structures, *Monatshefte f. Chemie*, 125:167–188, 1994.

[7] Krylov, A.S., Zasedateleva, O.A., Prokopenko, D.V., Rouviere-Yaniv, J., and Mirzabekov, A.D., Massive parallel analysis of the binding specificity of histone-like protein HU to single- and double-stranded DNA with generic oligodeoxyribonucleotide microchips, *Nucleic Acids Res.*, 12:2654–2660, 2001.

[8] Lindblad-Toh, K., Winchester, E., Daly, M.J., Wang, D.G., Hirschhorn, J.N., Laviolette, J.P., Ardlie, K., Reich, D.E., Robinson, E., Sklar, P., Shah, N., Thomas, D., Fan, J.B., Gingeras, T., Warrington, J., Patil, N., Hudson, T.J., and Lander, E.S., Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse, *Nat. Genet.*, 4:381–386, 2000.

[9] Lipman, D.J. and Pearson, W.R., Rapid and sensitive protein similarity searches, *Science*, 227:1435–1441, 1985.

[10] McCaskill, J.S., The equilibrium partition function and base pair binding probabilities for RNA secondary structure, *Biopolymers*, 29:1105–19, 1990.

[11] Pearson, W.R., Rapid and sensitive sequence comparison with FASTP and FASTA, *Methods Enzymol.*, 210:575–601, 1992.

[12] Pearson, W.R., Using the FASTA program to search protein and DNA sequence databases, *Computer Analysis of Sequence Data, Part I*, A.M. Griffin & H.G Griffin, Eds., Humana Press, Inc., Totowa, NJ, 25:307–331, 1994.

[13] Pearson, W.R. and Lipman, D.J., Improved tools for biological sequence comparison, *Proc. Natl. Acad. Sci. USA*, 85:2444–2448, 1988.

[14] Puglisi, J.D. and Tinoco Jr., I., Absorbance melting curves of RNA, *Methods in Enzymology*, 180:304–325, 1989.

[15] Schena, M., Shalon, D., Davis, R.W., and Brown, P.O., Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science*, 270(5235):467–470, 1995.

[16] Seetharaman, S., Zivarats, M., Sudarsan, N., and Braker, R.R., Immobilized RNA switches for the analysis of complex chemical and biological mixture, *Nature Biotechnology*, 19:336–341, 2001.

[17] Shoemaker, D.D., Schadt, E.E., Armour, C.D., He, Y.D., Garrett-Engele, P., McDonagh, P.D., and Loer, P.M., Experimental annotation of the human genome using microarray technology, *Nature*, 409:922–927, 2001.