DNA Solitons and Codon Bias

Alex Kasman
Department of Mathematics
College of Charleston

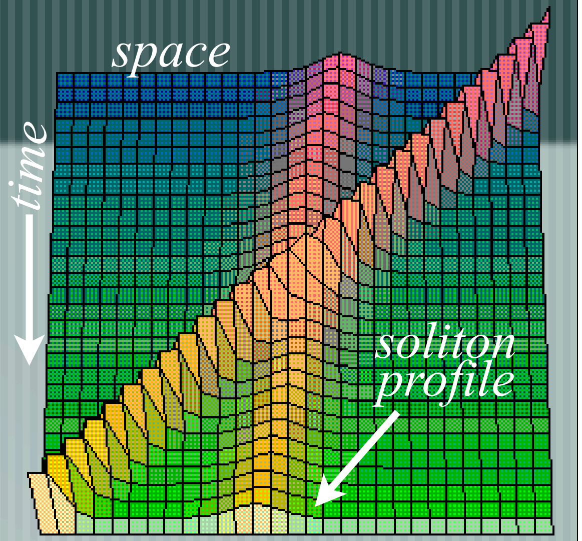
These are (very) preliminary results from a mathematician hoping to do more work in mathematical biology. Englander et al have modeled the transcription fork of DNA as a soliton of a discrete sine-Gordon model. Here I will propose that when the sequence is taken into account, such a model may help to explain the existence of introns and codon bias. Feedback and suggestions would be greatly appreciated.

KdV Solitons

- J.S. Russell observed a persistent, solitary wave on a canal in 1834. These are modeled by the KdV equation (1895), a nonlinear PDE.

- Essentially, all localized disturbances are treated by the KdV equation as an interaction of these particle-like humps. See [1,2].

- Amazingly, the existence of soliton solutions seems to be tied to integrability, allowing us to find exact solutions to nonlinear dynamical systems.



Note that in this "collision" of KdV solitons, they do not simply pass through each other, but actually divert the path slightly.

Sine-Gordon and DNA Models:

(1) $$\phi_{xx}(x,t) - \phi_{tt}(x,t) = \sin\left(\phi(x,t)\right)$$

(2) $$\phi(x,t) = 4\arctan\left(\exp\left[\frac{x - vt}{\sqrt{1 - v^2}}\right]\right)$$

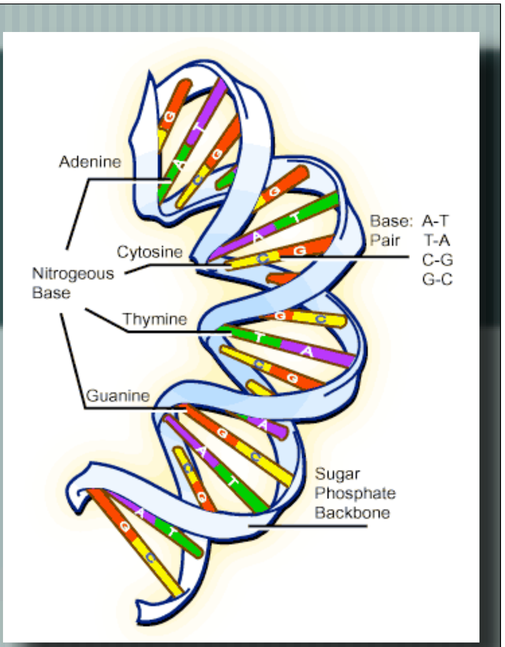
(3) $$\ddot{\phi}_n = (\phi_{n+1} - 2\phi_n - \phi_{n-1}) - \sin\phi_n$$

(4) $$\ddot{\phi}_n = (\phi_{n+1} - 2\phi_n + \phi_{n-1}) - \frac{2e_n}{5}\sin(\phi_n)$$

Physical model of (3). Note that the value of Φ is an angle so that 0 and 2π describe the same "rest state" and other values involve some "twist".

- Eqn (1) is the sine-Gordon equation used in particle physics.
- It has exact soliton solution (2) and also anti-soliton solutions.
- Englander modeled DNA dynamics using (3) where continuous space x has been replaced by discrete bases, n. See reference [3].
- Finally, Salerno [4] added n-dependent bond strength e_n to allow sequence to affect dynamics (4). If e_n=5/2 then (4) becomes (3). Biologically, $e_n \in \{2,3\}$.

Codon Bias and Introns

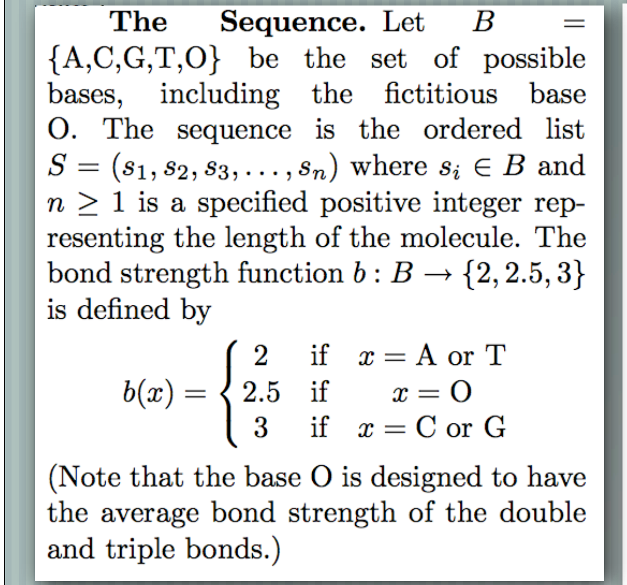


- The DNA code is made up of A,C,T and G where A and T have bond strength e_n = 2 and C or G have bond strength e_n = 3.

- DNA Code is <u>redundant but biased</u>: For instance, TTA and CTG both represent leucine. However, living cells demonstrate a bias...some mechanism seems to prevent random selection of codons. (See [5]-[9] in references.)

- Often, in the middle of a sequence encoding a protein, an "<u>intron</u>" appears which does not belong in that protein. The cell is able to snip out this unwanted piece, but why is it there in the first place? It could be non-optimal result of evolution, but nonlinear dynamics suggests a functional explanation.

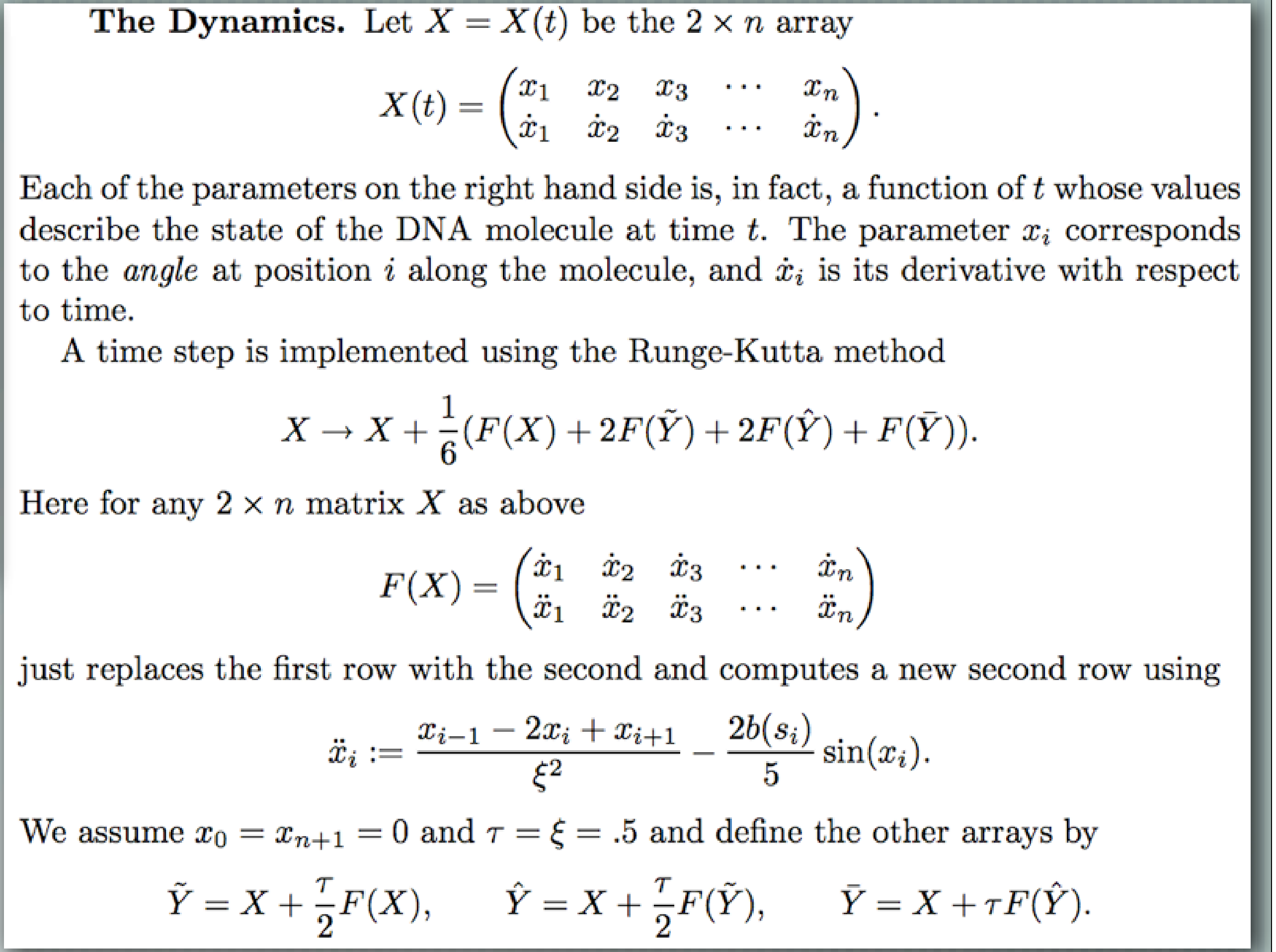
Numerical Integration



The Sequence. Let $B = \{$A,C,G,T,O$\}$ be the set of possible bases, including the fictitious base O. The sequence is the ordered list $S = (s_1, s_2, s_3, \ldots, s_n)$ where $s_i \in B$ and $n \geq 1$ is a specified positive integer representing the length of the molecule. The bond strength function $b : B \to \{2, 2.5, 3\}$ is defined by

$$b(x) = \begin{cases} 2 & \text{if} \quad x = \text{A or T} \\ 2.5 & \text{if} \quad x = \text{O} \\ 3 & \text{if} \quad x = \text{C or G} \end{cases}$$

(Note that the base O is designed to have the average bond strength of the double and triple bonds.)



The Dynamics. Let $X = X(t)$ be the $2 \times n$ array

$$X(t) = \begin{pmatrix} x_1 & x_2 & x_3 & \cdots & x_n \\ \dot{x}_1 & \dot{x}_2 & \dot{x}_3 & \cdots & \dot{x}_n \end{pmatrix}.$$

Each of the parameters on the right hand side is, in fact, a function of t whose values describe the state of the DNA molecule at time t. The parameter x_i corresponds to the *angle* at position i along the molecule, and $\dot{x}_i$ is its derivative with respect to time.

A time step is implemented using the Runge-Kutta method

$$X \to X + \frac{1}{6}(F(X) + 2F(\tilde{Y}) + 2F(\hat{Y}) + F(\bar{Y})).$$

Here for any $2 \times n$ matrix X as above

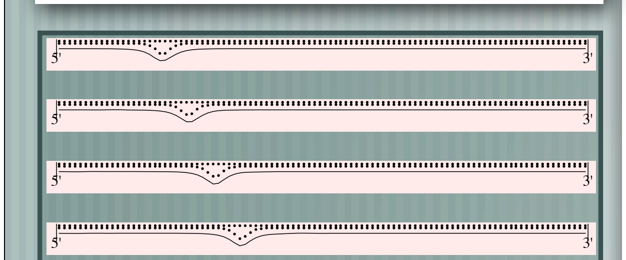
$$F(X) = \begin{pmatrix} \dot{x}_1 & \dot{x}_2 & \dot{x}_3 & \cdots & \dot{x}_n \\ \ddot{x}_1 & \ddot{x}_2 & \ddot{x}_3 & \cdots & \ddot{x}_n \end{pmatrix}$$

just replaces the first row with the second and computes a new second row using

$$\ddot{x}_i := \frac{x_{i-1} - 2x_i + x_{i+1}}{\xi^2} - \frac{2b(s_i)}{5}\sin(x_i).$$

We assume $x_0 = x_{n+1} = 0$ and $\tau = \xi = .5$ and define the other arrays by

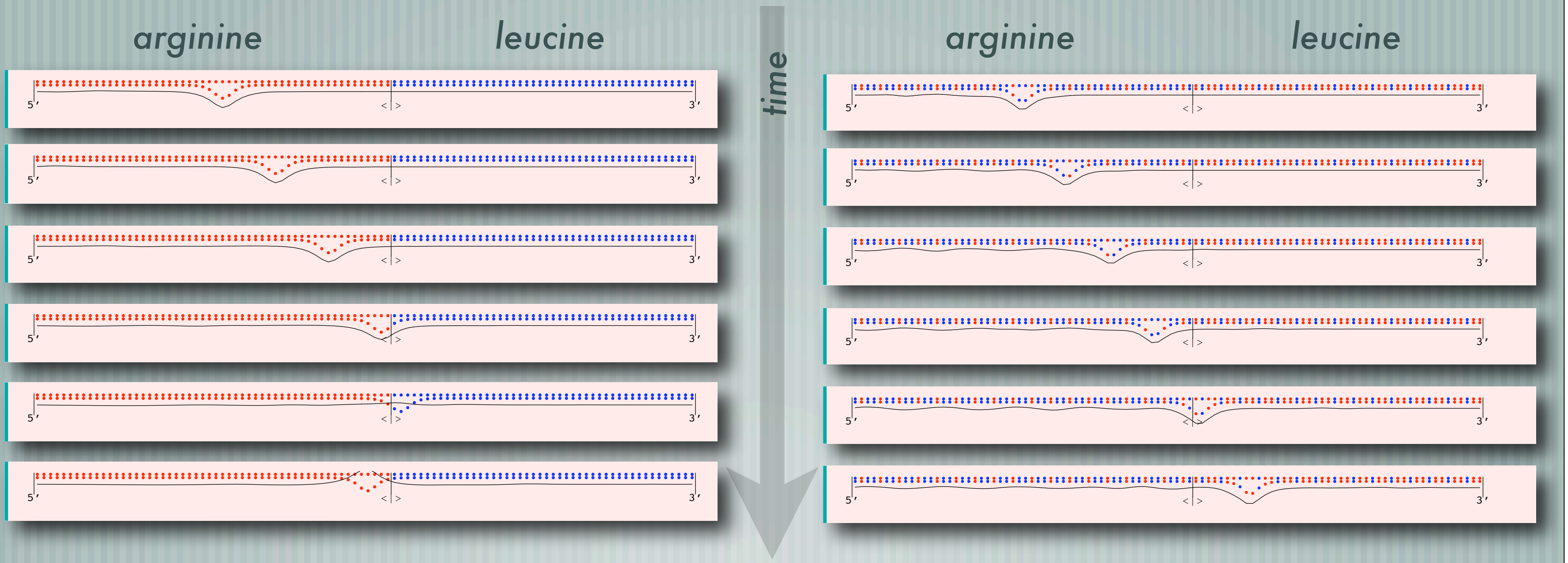
$$\tilde{Y} = X + \frac{\tau}{2}F(X), \qquad \hat{Y} = X + \frac{\tau}{2}F(\tilde{Y}), \qquad \bar{Y} = X + \tau F(\hat{Y}).$$



If S=(0,0,0,...) then the soliton just travels along the molecule smoothly.

Codon Selection Makes a Difference

The amino acid arginine has the codon CGC which is 333 as well as the codon AGA which is 232. Similarly, Leucine can be TTA (222) or CTG (323). If we select a sequence in which repeating low energy leucine is followed immediately by high bond strength argenine bonds the soliton is effectively stopped before completing the sequence, but if the other codon selections are used it works just fine:



They encode the same protein, but transcription fails on the left.

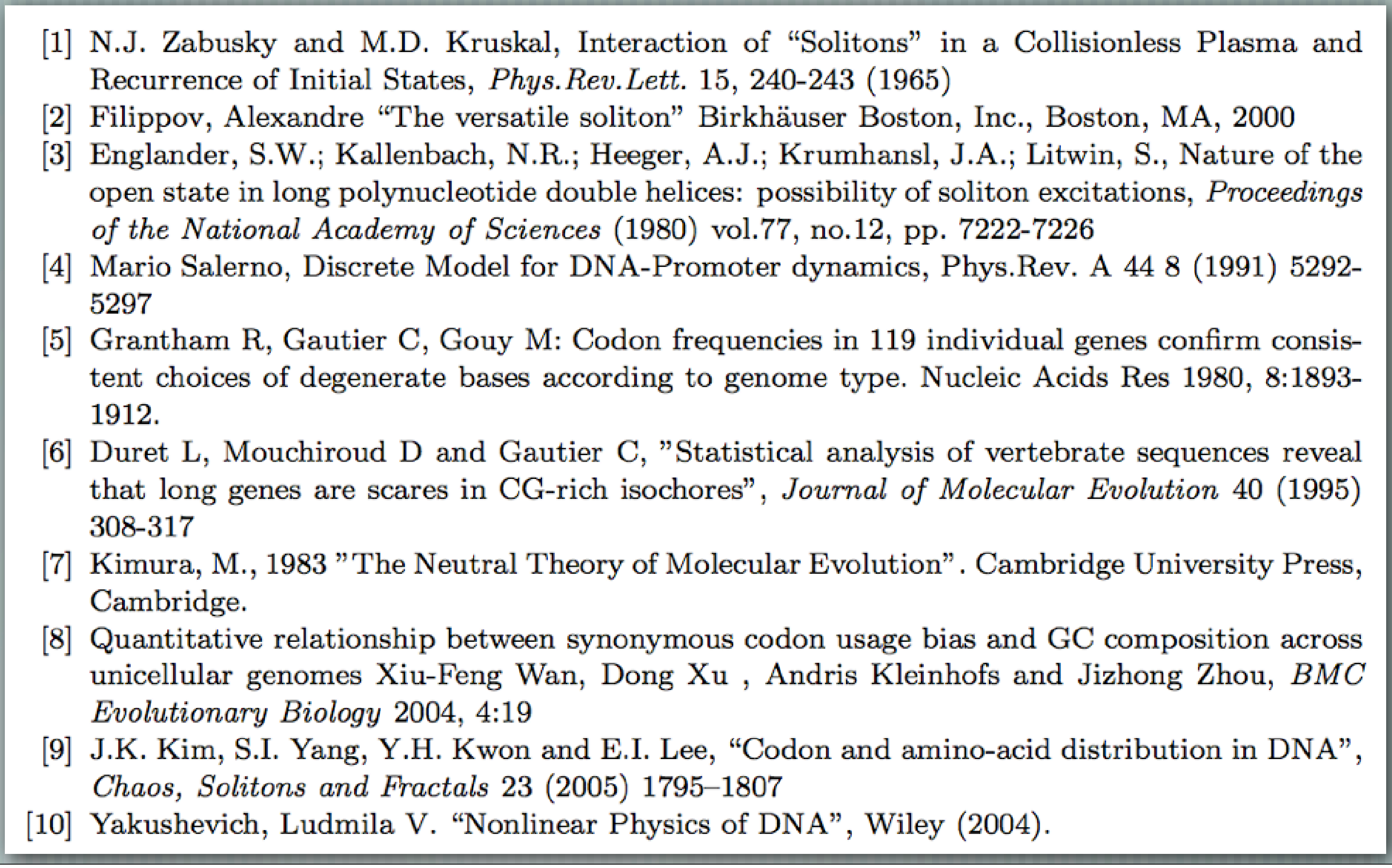
Future Plans

- Better *in silico* experiments, either using a more sophisticated numerical algorithm or a more realistic DNA model.
- Demonstrate that inserting an intron as a preconditioning sequence can allow soliton to continue past a “bad” segment.
- Search real sequence databases for statistical evidence.
- Perform *in vitro* experiments (with a biologist) to demonstrate that apparently equivalent sequences can be transcribed with different rates of efficiency.

Conclusions

Although the model and the numerics here are not very sophisticated, this still serves as a proof of concept. If nonlinear dynamics induced by the molecular attraction of the bases is involved, then the particular sequence may affect the efficiency (or possibility) of transcription. As we've seen here, this could provide a functional role for codon bias or introns not previously considered by biologists. (Then again, I could be completely wrong! Your advice, suggestions would be greatly appreciated.)

References



[1] N.J. Zabusky and M.D. Kruskal, Interaction of "Solitons" in a Collisionless Plasma and Recurrence of Initial States, *Phys.Rev.Lett.* 15, 240-243 (1965)

[2] Filippov, Alexandre "The versatile soliton" Birkhäuser Boston, Inc., Boston, MA, 2000

[3] Englander, S.W.; Kallenbach, N.R.; Heeger, A.J.; Krumhansl, J.A.; Litwin, S., Nature of the open state in long polynucleotide double helices: possibility of soliton excitations, *Proceedings of the National Academy of Sciences* (1980) vol.77, no.12, pp. 7222-7226

[4] Mario Salerno, Discrete Model for DNA-Promoter dynamics, Phys.Rev. A 44 8 (1991) 5292-5297

[5] Grantham R, Gautier C, Gouy M: Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. Nucleic Acids Res 1980, 8:1893-1912.

[6] Duret L, Mouchiroud D and Gautier C, "Statistical analysis of vertebrate sequences reveal that long genes are scares in CG-rich isochores", *Journal of Molecular Evolution* 40 (1995) 308-317

[7] Kimura, M., 1983 "The Neutral Theory of Molecular Evolution". Cambridge University Press, Cambridge.

[8] Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes Xiu-Feng Wan, Dong Xu , Andris Kleinhofs and Jizhong Zhou, *BMC Evolutionary Biology* 2004, 4:19

[9] J.K. Kim, S.I. Yang, Y.H. Kwon and E.I. Lee, "Codon and amino-acid distribution in DNA", *Chaos, Solitons and Fractals* 23 (2005) 1795–1807

[10] Yakushevich, Ludmila V. "Nonlinear Physics of DNA", Wiley (2004).