Biopolymer Chain Elasticity with *Mathematica*

a novel concept to fold DNA or RNA into 3-D hairpin structures

Guillaume P. H. Santini

Laboratoire de Biophysique Moléculaire, Cellulaire et Tissulaire (BioMoCeTi), UMR CNRS 7033
Universités Pierre et Marie Curie Paris6 et Paris13, Genopole Campus 1, RN7, 91030 Evry, France

Christophe Pakleza

Epigenomics Project, Genopole, 91030 Evry, France

Jean A. H. Cognet

Laboratoire de Biophysique Moléculaire, Cellulaire et Tissulaire (BioMoCeTi), UMR CNRS 7033
Universités Pierre et Marie Curie Paris6 et Paris13, Genopole Campus 1, RN7, 91030 Evry, France

`cognet@ccr.jussieu.fr`

Abstract of "Talk"

The tri-dimensional structures of DNA or RNA hairpin molecules exhibit a considerable variety of complex shapes. They constitute, with double helices, one of the most fundamental structural units of DNA or RNA. They are formed from a single strand molecule and consist of a base-paired helical stem structure and of a closing loop sequence with unpaired or mismatched nucleotides.

We propose a new molecular modelling methodology, called the Biopolymer Chain Elasticity (BCE) approach. It is based on the theory of elasticity of thin rods and on the assumption that single-stranded B-DNA or A-RNA behaves as a continuous, unshearable, unstretchable and flexible thin rod. This approach is implemented in a computer program, S-mol©, written in *Mathematica* to fold short single stranded oligonucleotides into hairpins. S-mol is designed as a general laboratory bench for doing molecular modelling of biological macromolecules. It takes advantage of the main features of *Mathematica* to develop a full command language, to represent molecules both as discrete (atoms) and continuous objects (3D trajectories), to compute different geometrical description parameters, and to perform global deformations of these large and complex objects without loosing the local geometries at the atomic level.

BCE theoretical approach and S-mol are capable of reproducing the tri-dimensional course of the hairpin loop chain and, using experimental data, of reproducing models very close to published solution structures. The geometry of end conditions imposed by the double helical stem is sufficient to dictate the different characteristic DNA or RNA loop shapes. The natural loop description provided by the elastic line model and by the new independent parameter, Ω, which corresponds to the rotation angle of nucleosides about the elastic line, offers a simple, global, coherent, and quantitative description of biological hairpin molecules.

Background

Hairpin structures constitute, with double helices, one of the most fundamental structural units of DNA or RNA. They are formed from a single strand molecule and consist of a base-paired stem structure and of a loop sequence with unpaired or mismatched nucleotides as shown in Figure 1D. RNA hairpins have been known to play essential structural and biological roles for several decades. Biological importance of DNA hairpins has long been suspected but has been demonstrated only recently (for a brief review see Pakleza et Cognet, 2003). *In vivo*, DNA hairpins occur as transient single-stranded intermediates in many aspects of DNA metabolism including DNA replication, repair, and recombination and can exist when intrastrand pairing occurs between inverted repeats. The recent developments and successes of the aptamer strategy and the use of non-pathogenic viruses as a gene therapy vector have provided further motivation for understanding DNA hairpin structures. Hairpin-like structural elements are essential both for replication of the virus and for site-specific integration.

Numerous thermodynamical and structural studies of DNA and RNA hairpins have been reported (for reviews see Hilbers *et al.*, 1994; Varani, 1995). At present, a large number of solution structures of DNA or RNA hairpins, on the order of a hundred are available and a considerable wealth of variations is observed in these tri-dimensional structures. We have focused our attention on molecular structures of well-characterised stable hairpins, with or without base pairing in the loop, with different lengths and with DNA or RNA sequences. The list, given in Table 1, includes all known solution structures of hairpins with -TTT- in the loop. For these molecules, there are no base pairing in the loop and, for most of them NMR distances are available. The eight remaining molecules in Table 1 correspond to the structures of eight remarkably stable DNA or RNA hairpin molecules closed by a mispair, recently determined in solution by NMR and deposited in the PDB. The sequence are: one DNA tetraloop, -GTTA-; three DNA triloops, -AAA- or -GCA-; and four RNA tetraloops, -UUCG-. We postulate in agreement with reports on DNA and RNA hairpin structures of Table 1, that the sugar phosphate backbone of the stem continues in the loop as a right-handed B-DNA or A-RNA helix and that most torsion angles in the loop are close to classical B-DNA or A-RNA values (Pakleza et Cognet, 2003; Santini *et al.*, 2003). These observations suggest that hairpin structures might be obtained by bending as smoothly as possible a single strand B-DNA or A-RNA helix into a hairpin fold.

Methods and Results

We have developed a new molecular modelling methodology and a computer program, S-mol© (Pakleza et Cognet, 2003; Santini *et al.*, 2003), based on the theory of elasticity of thin rods to fold short single stranded oligonucleotides into hairpins (Landau and Lifshitz, 1970; Shi and Hearst, 1994; Tobias *et al.*, 1994). It is called the Biopolymer Chain Elasticity (BCE) approach and is used to predict the tri-dimensional backbones of DNA and RNA hairpin loops. Note that elasticity theory of thin rods has been successfully applied to DNA in the different context of long double stranded helical DNA in the range of hundreds of base pairs. The latter approach should not be confused with our methodology that was applied for the first time at the level of several nucleotides.

The BCE approach is based on the theory of elasticity of thin rods and on the assumption that single-stranded B-DNA or A-RNA behaves as a continuous, unshearable, unstretchable and flexible thin rod. The sugar phosphate chain is modelled by different helical segments as shown in Figure 1 A-B. The BCE approach requires four construction steps illustrated in Figure 1. 1/ Computation of the tri-dimensional trajectory of the folded elastic line in Figure 1C according to the prescribed boundary conditions of Figure 1 A-B (matching of the locations and directions of the tangents shown as green and red arrows). Note that from a physical point of view, the trajectory of the loop is defined by its length, by its elastic properties, and by applied forces and moments at extremities of the rod. However from a mathematical point of view, the whole trajectory can be deduced from its length, and from

the sole geometry of end conditions. 2/ Global deformation of single-stranded helical DNA or RNA molecular structure onto the elastic line by a geometrical transformation (Figure 1D). The overall molecular architecture of the sugar phosphate backbone is defined at this step. 3/ Optimization of the nucleoside rotation angles $\Delta\Omega$ about the elastic line (top part of Figure 1D). The two previous modelling steps provide the overall geometry of the backbone and a set of local (Frenet) reference frames on the elastic line, which can be used to define the angle to rotate blocks of atoms about their closest tangents on the elastic line. These rotations are necessary to reproduce the conformations of nucleotides observed experimentally. 4/ Energy minimization to restore backbone bond lengths and bond angles (not shown).

S-mol is designed as a general laboratory bench for doing molecular modelling of biological macromolecules. It takes advantage of the main features of *Mathematica* to develop a full command language, to represent molecules both as discrete (atoms) and continuous objects (3D trajectories), to compute different geometrical description parameters, and to perform global deformations of these large and complex objects without loosing the local geometries at the atomic level.

S-mol can generate single or double-stranded helical DNA or RNA. Alternatively it can read a file that contains a description of the biological macromolecule under study. This description consists of the list of names of all atoms with their cartesian coordinates, the names of the residues to which they belong and the residue number within the molecular chain. It is read from a file in the standard Brookhaven Protein DataBank format (PDB) or in the AMBER format (Case *et al.*, 2005). The main capabilities of S-mol are summarized in Figure 2. They include many standard molecular modelling functions (editing, visualisation, output...), as well as functions dedicated to nucleic acids analyses (torsion angles analysis, sugar puckering, ...).

Three different methods (superpositions, distance of main chain atoms to the elastic line, RMSd of NMR derived distances and/or RMSd between molecules) were used to show a very good agreement between the trajectories of sugar phosphate backbones and, between entire molecules of theoretical models and of published solutions conformations. As a result, with this simple idea, we have shown that single-stranded B-DNA can be deformed into hairpin loops that match not only all published NMR data available for trinucleotide TTT loops (Pakleza et Cognet, 2003) but also the PDB structures of tri- and tetra-loops of DNA (Santini *et al.*, 2003). We have shown in addition that single-stranded A-RNA can be deformed with the same folding methodology into UUCG tetraloops (Santini et al., 2003). Note the shapes of DNA and RNA hairpins are different (cf. Figure 1E), but are well reproduced by the same methodology applied with the different end conditions imposed by B-DNA or A-RNA helical geometries. Note also that both types of hairpins structures without and with base pairing in the loop are well reproduced by the same BCE methodology.

Conclusion

Single-stranded B-DNA and A-RNA behave in these case studies to first approximation as a continuous, unstretchable and flexible thin rod where most torsion angle values are preserved in the folding process. This "flexibility" is taken in the sense of the elasticity theory of thin rods and in the sense of the least deformation energy principle. Our theoretical approach is a molecular modelling methodology capable of predicting the tri-dimensional course of the sugar phosphate chain from boundary conditions and, using NMR derived distances or structure constraints, of generating models very close to published solution structures. The natural description of loop folding with the new parameter angles, Ω, offers a considerable simplification of the molecular modelling of hairpin loops and a reduction in number of conformation parameters. Furthermore Ω angles can be varied independently from each other, since the global shape of the hairpin loop is preserved in all cases.

More studies are needed to check whether other hairpins can be reproduced and described with the BCE approach. If so, they would provide the first quantitative measurements to classify and to understand the structures of DNA and RNA hairpin loops and possibly of many other important biological macromolecules.

References

Allain,F.H.-T., Howe,P.W.A., Neuhaus,D. & Varani,G. (1997). Structural basis of the RNA-binding specificity of human U1A protein. *EMBO J.*, **16**, 5764-5774.

Allain,F.H.-T. & Varani,G. (1995). Structure of the P1 helix from Group I Self-splicing introns. *J. Mol. Biol.*, **250**, 333-353.

Boulard,Y., Gabarro-Arpa,J., Cognet,J.A.H., Le Bret,M., Guy,A., Téoule,R., Guschlbauer,W. & Fazakerley,G.V. (1991). The solution structure of a DNA hairpin containing a loop of three thymidines determined by nuclear magnetic resonance and molecular mechanics. *Nucleic Acids Res.*, **19**, 5159-5167.

Butcher,S.E., Allain,F.H.-T. & Feigon,J. (1999). Solution structure of the B domain from the hairpin ribozyme. *Nature struct. Biol.*, **6**, 212-216.

Case,D.A. Cheatham,T.E.III, Darden,T., Gohlke,H., Luo,R. Merz,K.M., Onufriev,A., Simmerling,C. Wang,B., Woods,R.J. (2005) The Amber Biomolecular Simulation
Programs. *J. Comp. Chem.*, **26**, 1668-1688.

Chou,S.H., Tseng,Y.Y. & Chu,B.Y. (2000). Natural abundance heteronuclear NMR studies of the T3 mini-loop hairpin in the terminal repeat of the adenoassociated virus 2. *J. Biomol. NMR*, **17**, 1-16.

Chou,S.-H., Zhu,L., Gao,Z., Cheng,J.-W. & Reid,B.R. (1996). Hairpin Loops Consisting of Single Adenine Residues Closed by Sheared A.A and G.G pairs Formed by the DNA Triplets AAA and GAG: Solution Structure of the d(GTACAAAGTAC) Hairpin. *J. Mol. Biol.*, **264**, 981-1001.

Colmenarejo,G. & Tinoco,I.Jr (1999). Structure and Thermodynamics of Metal Binding in the P5 helix of a Group I Intron Ribozyme. *J. Mol. Biol.*, **290**, 119-135.

Geomview, The Geometry Center, University of Minnoseta, Minneapolis, USA.

Hilbers,C.W., Heus,H.A., van Dongen,M.J. & Wijmenga,S.S. (1994). The hairpin elements of nucleic acid structure: DNA and RNA folding. In Nucleic Acids and Molecular Biology (Eckstein, F. & Lilley, D. M. J., eds.), pp. 56-104. Springer Verlag, Berlin.

Kuklenyik,Z., Yao,S. & Marzilli,L.G. (1996). Similar conformations of hairpins with TTT and TTTT sequences: NMR and molecular modeling evidence for T.T base pairs in the TTTT hairpin. *Eur. J. Biochem.*, **236**, 960-969.

Landau,L.D. and Lifshitz,E.M. (1970). Theory of Elasticity, pp. 165 3rd ed., Pergamon, Oxford.

Mooren,M.M.W., Pulleyblank,D.E., Wijmenga,S.S., van de Ven,F.J. & Hilbers,C.W. (1994). The solution structure of the hairpin formed by d(TCTCTC-TTT-GAGAGA). *Biochemistry*, **33**, 7315-7325.

Pakleza,C., and Cognet,J.A.H. (2003). Biopolymer Chain Elasticity: a novel concept and a least deformation energy principle predicts backbone and overall folding of DNA TTT hairpins in agreement with NMR distances. *Nucleic Acids Res.*, **31**, 1075-1086.

Santini,G.P.H., Pakleza,C., and Cognet,J.A.H. (2003) "DNA tri- and tetra-Loops and RNA tetra-Loops hairpins fold as Elastic Biopolymer Chains in agreement with PDB coordinates", *Nucleic Acids Res.*, **31**, 1086-1096.

Shi,Y. & Hearst,J.E. (1994). The Kirchhoff elastic rod, the nonlinear Schrö   dingequation, and DNA supercoiling. *J. Chem. Phys.*, **101**, 5186-5200.

Tobias,I., Coleman,B.D. & Olson,W.K. (1994). The dependence of DNA tertiary structure on end conditions: Theory and implications for topological transitions. *J. Chem. Phys.*, **101**, 10990-10996.

van Dongen,M.J.P., Mooren,M.M.W., Willems,E.F.A., van der Marel,G.A., van Boom,J.H., Wijmenga,S.S. & Hilbers,C.W. (1997). Structural features of the DNA hairpin d(ATCCTA-GTTA-TAGGAT): formation of a G-A pair in the loop. *Nucleic Acids Res.*, **25**, 1537-1547.

Varani,G. (1995). Exceptionally stable nucleic acid hairpins. *Annu. Rev. Biophys. Biomol. Struct.*, **24**, 379-404.

WolframResearch, Inc. (1999) *Mathematica*, Version 4.0, Champaingn, IL, USA.

Zhu,L., Chou,S.-H. & Reid,B.R. (1996). A single G-to-C change causes human centromere TGGAA repeats to fold

back into hairpins. ***Proc. Natl. Acad. Sci. USA***, **93**, 12159-12164.

Zhu,L., Chou,S.-H., Xu,J. & Reid,B.R. (1995). Structure of a single-cytidine hairpin loop formed by the DNA triplet GCA. ***Nature Struct. Biol.***, **2**, 1012-1017.

Acknowledgements

Figures

Figure 1.

Schematic overview of the construction process of an RNA tetraloop hairpin using the Biopolymer Chain Elasticity approach: (A) a continuous and flexible thin rod, represented by a ribbon for better visibility is associated to a four nucleotides helical segment; (B) a double helical RNA segment is generated along helical lines; (C) the flexible rod is bent into the elastic solution curve so that the tangents at its extremities, shown as green and red arrows, match those of the two helices; (D) the complete molecular structure is computed after global deformation. (E) Superimposed views into the minor groove of the computed elastic rod curve for a DNA tetraloop hairpin, shown in red, and for an RNA tetraloop hairpin shown in blue. The radii of cylinders and circles are those of the sugar phosphate backbones. Top part displays schematic views of nucleoside block rotations with angle, $\Delta\Omega$ about the elastic rod curve.
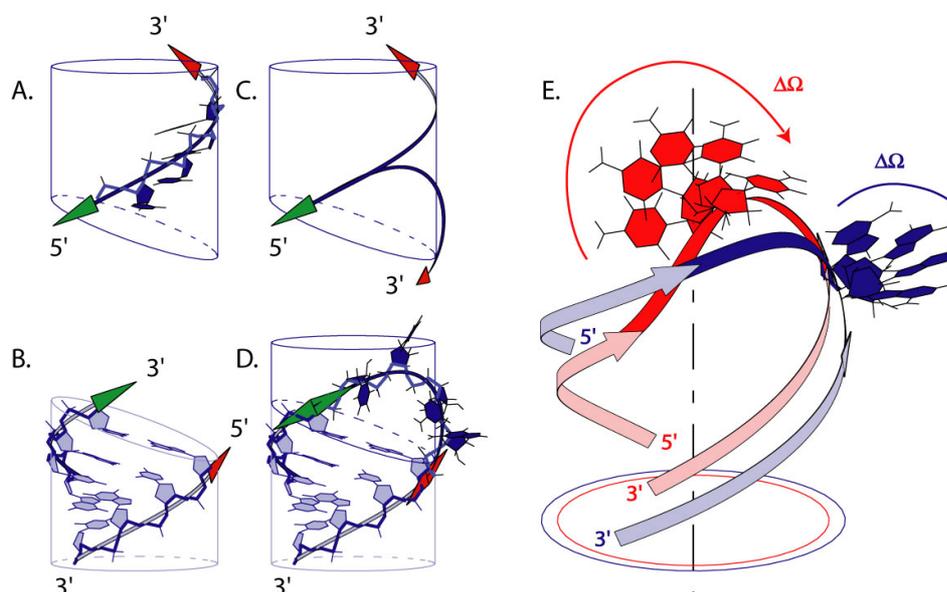
Figure 2.

Schematic overview of the S-mol program: (Top) starting from the left hand side in a clockwise direction, a helical DNA or RNA molecule can be generated within S-mol, or a biological macromolecule can be read from a Protein DataBank file; the trajectory of a continuous and flexible thin rod is computed using the theory of elasticity for the loop fragment; the complete molecular structure is computed after global deformation as explained in Figure 1. (Bottom) In any suitable order, the biological

macromolecules can be edited and modified, written in different file formats (PDB, AMBER or S-mol), visualized as geometric objects with Geomview, compared to different molecules with superimposition tools, and analyzed.
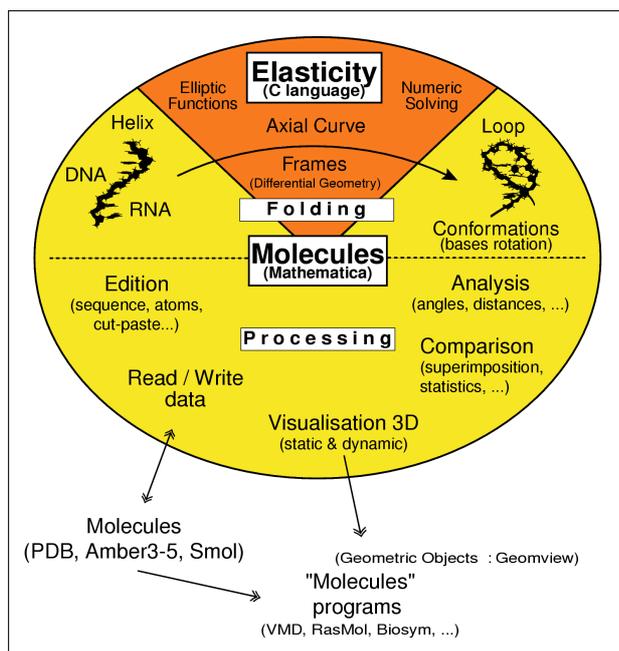
Table

Table 1:

Molecular structures of published hairpins well reproduced by the BCE approach, with PDB identification, original authors, and DNA or RNA sequences. The first four (*) were analysed using NMR derived distances and NMR derived solution structures (Pakleza et Cognet, 2003). The last eight (**) are selected from the Protein Data Bank and were analysed by an automated procedure (Santini *et al.*, 2003).

PDB id	Authors	DNA or RNA Sequence (experiments)
DNA	**NMR derived distances and structures**	**No pairing in the loop**
	Boulard *et al.*,1991∗	d(tctctc−T1T2T3−gagaga)
	Mooren *et al.*,1994∗	d(tctctc−T1T2T3−gagaga)
	Kuklenyik *et al.*,1996∗	d(gcgc−T1T2T3−gcgc)
	Chou *et al.*,2000∗	d(gaagc−T1T2T3−gcttc)
DNA	**Sets of PDB conformations**	**With pairing G.A or A.A in the loop**
1ac7	van Dongen *et al.*,1997∗∗	d(...ccta−G1T2T3A4−tagg...)
1bjh	Chou *et al.*,1996∗∗	d(gtac−A1A2A3−gtac)
1xue	Zhu *et al.*,1996∗∗	d(...gaat−G1C2A3−atgg...)
1zhu	Zhu *et al.*,1995∗∗	d(caat−G1C2A3−atg)
RNA	**Sets of PDB conformationss**	**With pairing U.G in the loop**
1aud	Allain *et al.*,1997∗∗	r(...gucc−U1U2C3G4−ggac...)
1b36	Butcher *et al.*,1999∗∗	r(...gcgc−U1U2C3G4−gcgc...)
1c0o	Colmenarejo&Tinoco,1999∗∗	r(...gguc−U1U2C3G4−gguc...)
1hlx	Allain&Varani,1995∗∗	r(...uaac−U1U2C3G4−guug...)