BpForms: a toolkit for concretely describing modified DNA, RNA and proteins

Paul F. Lang,[†,‡,¶] Yassmine Chebaro,[†,‡,§] and Jonathan R. Karr[*,†]

†*Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA.*

‡*Contributed equally to this work*

¶*Department of Biochemistry, Oxford University, South Parks Road, Oxford OX1 3QU, UK.*

§*Institut de Génétique et de Biologie Moléculaire et Cellulaire, Institut National de la Santé et de la Recherche Médicale, Centre National de la Recherche Scientifique, Université de Strasbourg, Illkirch, France.*

E-mail: karr@mssm.edu

Abstract

Summary: Non-canonical nucleic and amino acid monomers are essential to enhance the information capacity, functional capabilities, and stability of DNA, RNA, and protein biopolymers. However, there are few tools for describing the primary structure of biopolymers that include non-canonical monomers. We developed *BpForms*, the first toolkit for concretely and compactly describing the primary structures of non-canonical 1-dimensional biopolymers. *BpForms* includes the first alphabets of non-canonical DNA, RNA, and protein monomers; a FASTA-like notation for describing biopolymers; and a website, a command line program, a REST API, and a Python package for calculating properties of biopolymers. We anticipate *BpForms* will be a

valuable tool for communicating data about modified DNA, RNA, and proteins, as well as integrating data about epigenetic, post-transcriptional, and post-translational modification. *BpForms* will also be valuable for whole-cell modeling and cell engineering.

Availability and implementation: *BpForms* is freely available open-source at https://bpforms.org.

Introduction

Non-canonical nucleic and amino acid monomers are essential to enhance the information capacity, functional capabilities, and stability of DNA, RNA, and protein biopolymers. For example, methylation helps bacteria distinguish self from foreign DNA and pseudouridine helps tRNA complement multiple codons. The FASTA format describes the structures of 1-dimensional biopolymers composed of canonical monomers. Recently, the Consortium for Top Down Proteomics developed the ProForma format to describe modified proteins[4]. However, ProForma is limited to proteins composed of monomers in databases such as RESID[2], it is verbose due to the lack of an expanded alphabet, and there is no public software implementation of ProForma. We developed *BpForms*, the first toolkit for concretely and compactly describing the primary structures of non-canonical biopolymers and computing their properties. *BpForms* uses alphabets and a FASTA-like notation to concretely describe biopolymers, including the linkages between successive monomers and uncertainties in their structures; *BpForms* includes the first three alphabets of non-canonical DNA, RNA, and protein monomers; and *BpForms* includes a website, a command line interface (CLI), a REST API, and a Python API for calculating properties of biopolymers (Fig. 1). We developed *BpForms* to concretely describe the biopolymers represented by whole-cell (WC) computational models[3]. For the first time, *BpForms* will also enable concrete communication about non-canonical biopolymers. In turn, we anticipate that *BpForms* will facilitate data integration about epigenetic, post-transcriptional, and post-translational modification;

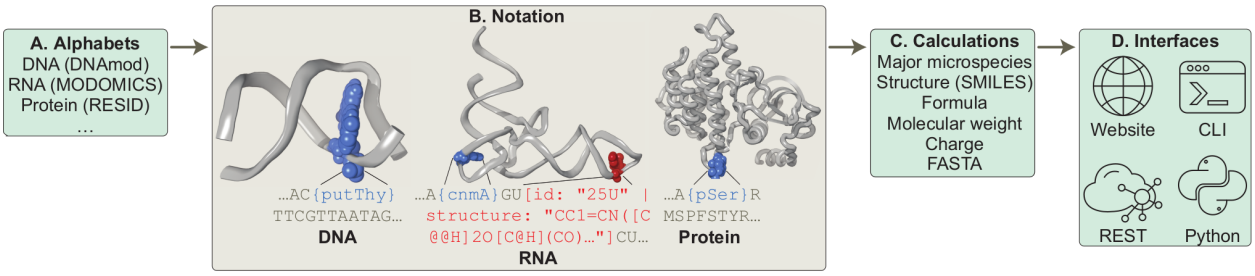

Figure 1: The *BpForms* toolkit helps researchers communicate and integrate data about non-canonical DNA, RNA, and protein biopolymers. The *BpForms* notation (**B**) concretely and compactly describes biopolymers as sequences of canonical (grey) and non-canonical monomers defined in an alphabet (blue) (**A**) or defined "inline" (red). This enables the *BpForms* software to calculate properties of biopolymers such as their formula (**C**). These calculations are available through the *BpForms* website, CLI, REST API, and Python package (**D**).

reconstruction of the structures of modified biopolymers and the reactions which produce them; and the design of expanded genetic codes. Here, we describe the *BpForms* alphabets, notation, and software tools.

Representation of non-canonical biopolymers

BpForms represents linear and circular 1-dimensional biopolymers as sequences of monomers (e.g., nucleobases) connected via backbones (e.g., deoxyribose 5-phosphates) and bonding operations (e.g., phosphodiester bond formation).

Alphabets of non-canonical monomers

BpForms uses alphabets to compactly describe biopolymers. Each alphabet is a list of the codes and structures of monomers. *BpForms* includes three alphabets of DNA nucleobases, RNA nucleosides, and protein residues curated from DNAmod[5], MODOMICS[1], and RESID (Fig. 1A). For consistency with these resources, the DNA, RNA, and protein alphabets are alphabets of DNA nucleobases, RNA nucleosides, and protein residues, respectively. To create consistent alphabets for calculating properties of biopolymers, we excluded inconsis-

tent entries from these resources such as DNA nucleosides, RNA nucleotides, and protein biopolymers. We also excluded entries which do not have defined structures. Users can also extend these alphabets to incorporate newly discovered monomers, as well as create additional alphabets for other types of biopolymers such as actin filaments. As described in the documentation, we welcome contributions to the alphabets via Git pull requests.

Notation for non-canonical biopolymers

BpForms uses a FASTA-like format to concretely and compactly describe biopolymers as sequences of monomers of an alphabet (Fig. 1B). Monomers with single-character codes are denoted by their codes (e.g., 'A'). Monomers with multi-character codes are denoted by enclosing their codes in curly brackets (e.g., '{m2A}'). Monomers can also be defined "inline" by enclosing one or more attributes, separated by vertical pipes, inside square brackets (e.g., '[id: "m2C" — structure: "O=C1N..."]'). The *structure* attribute indicates the structure in SMILES format. The *monomer-bond-atom*, *monomer-displaced-atom*, *left-bond-atom*, *left-displaced-atom*, *right-bond-atom*, and *right-displaced-atom* attributes describe the linkages between the monomer and the backbone and between successive backbones. The *base-monomer* attribute indicates the immediate precursor to the monomer (e.g., the precursor of m2A is A). The *id* and other attributes capture metadata. Inline monomers can also describe uncertainty in the structure and location of monomers. The *delta-mass* and *delta-charge* attributes indicate additional mass and charge that have been observed, but cannot be interpreted. The *position* attribute describes uncertainty in the locations of the monomer. The inline *structure* and linkage attributes are required to calculate properties of biopolymers. All other inline attributes are optional.

Calculated properties of biopolymers

By concretely representing biopolymers, *BpForms* can calculate properties such as their major protonation and tautomerization states, structures (e.g. SMILES), formula, molecular weight, and charge at specific pHs (Fig. 1C).

Export to FASTA format

To facilitate interpretation of biopolymers and backward compatibility, *BpForms* can export biopolymers to FASTA. This uses the *base-monomer* attributes of monomers and the alphabet codes of their roots.

User interfaces

BpForms includes four user interfaces for calculating properties of biopolymers: a website ((https://bpforms.org), a REST API ((https://bpforms.org/api), and a CLI and Python API (hrefhttps://pypi.python.org/pypi/bpformshttps://pypi.python.org/pypi/bpforms) (Fig. 1D).

Tutorial and documentation

The website, REST API, and CLI contain inline instructions. A tutorial and documentation for the Python API are available at https://sandbox.karrlab.org and https://bpforms.rtfd.io.

Implementation

We implemented *BpForms* in Python using the Cement, ChemAxon Marvin, Flask-RESTPlus, Open Babel, and Zurb Foundation packages.

Conclusion

By concretely describing the primary structure of non-canonical biopolymers, we anticipate that *BpForms* will help researchers concretely communicate data about non-canonical biopolymers and the biochemical processes which generate them. In turn, this will help us and others develop integrated models of the synthesis, degradation, interactions, and roles of biopolymers. We also anticipate that *BpForms* will be a valuable tool for communicating synthetic genetic codes and proteins.

Funding

This work was supported by the National Institutes of Health [grant number R35 GM119771 and grant number P41 EB023912], the National Science Foundation [grant number 1649014], and the Engineering and Physical Sciences Research Council [grant number EP/L016494/1]. *Conflict of Interest:* none declared.

References

(1) Boccaletto, P., Machnicka, M. A., Purta, E., Piątkowski, P., Bagiński, B., Wirecki, T. K., de Crécy-Lagard, V., Ross, R., Limbach, P. A., Kotter, A., *et al.* (2017). MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic Acids Res*, **46**(D1), D303–D307.

(2) Garavelli, J. S. (2004). The RESID Database of Protein Modifications as a resource and annotation tool. *Proteomics*, **4**(6), 1527–1533.

(3) Goldberg, A. P., Szigeti, B., Chew, Y. H., Sekar, J. A., Roth, Y. D., and Karr, J. R. (2018). Emerging whole-cell modeling principles and methods. *Curr Opin Biotechnol*, **51**, 97–102.

(4) LeDuc, R. D., Schwämmle, V., Shortreed, M. R., Cesnik, A. J., Solntsev, S. K., Shaw, J. B., Martin, M. J., Vizcaino, J. A., Alpi, E., Danis, P., *et al.* (2018). ProForma: a standard proteoform notation. *J Proteome Res*, **17**(3), 1321–1325.

(5) Sood, A. J., Viner, C., and Hoffman, M. M. (2018). DNAmod: the DNA modification database. *bioRxiv*, page 071712.