BMC Bioinformatics

BioMed Central
The Open Access Publisher

Single-molecule dataset (SMD): a generalized storage format for raw and processed single-molecule data

BMC Bioinformatics

doi:10.1186/s12859-014-0429-4

Max Greenfeld (max.greenfeld@gmail.com)
Jan-Willem van de Meent (janwillem.vandemeent@gmail.com)
Dmitri S Pavlichin (dmitrip@stanford.edu)
Hideo Mabuchi (hmabuchi@stanford.edu)
Chris H Wiggins (chris.wiggins@columbia.edu)
Ruben L Gonzalez (rlg2118@columbia.edu)
Daniel Herschlag (herschla@stanford.edu)

Published online: 16 January 2015

ISSN 1471-2105

Article type Software

Submission date 14 May 2014

Acceptance date 11 December 2014

Article URL http://dx.doi.org/10.1186/s12859-014-0429-4

Single-molecule dataset (SMD): a generalized storage format for raw and processed single-molecule data

Max Greenfeld[1,2,†]
Email: max.greenfeld@gmail.com

Jan-Willem van de Meent[3,†]
Email: janwillem.vandemeent@gmail.com

Dmitri S Pavlichin[4]
Email: dmitrip@stanford.edu

Hideo Mabuchi[5]
Email: hmabuchi@stanford.edu

Chris H Wiggins[6]
Email: chris.wiggins@columbia.edu

Ruben L Gonzalez Jr[7*]
[*] Corresponding author
Email: rlg2118@columbia.edu

Daniel Herschlag[1,2,8,*]
Email: herschla@stanford.edu

[1] Departments of Chemical Engineering, Stanford University, Stanford, CA 94305, USA

[2] Departments of Biochemistry, Stanford University, Stanford, CA 94305, USA

[3] Departments of Statistics, Columbia University, New York, NY 10027, USA

[4] Departments of Physics, Stanford University, Stanford, CA 94305, USA

[5] Departments of Applied Physics, Stanford University, Stanford, CA 94305, USA

[6] Departments of Applied Physics and Applied Mathematics, Columbia University, New York, NY 10027, USA

[7] Departments of Chemistry, Columbia University, New York, NY 10027-3126, USA

[8] Dept. of Biochemistry, B400, Stanford University, Stanford, CA 94305-5307, USA

[*] Corresponding author. Dept. of Biochemistry, B400, Stanford University, Stanford, CA 94305-5307, USA

[†] Equal contributors.

Abstract

Background

Single-molecule techniques have emerged as incisive approaches for addressing a wide range of questions arising in contemporary biological research [1-4]. The analysis and interpretation of raw single-molecule data benefits greatly from the ongoing development of sophisticated statistical analysis tools that enable accurate inference at the low signal-to-noise ratios frequently associated with these measurements. While a number of groups have released analysis toolkits as open source software [5-14], it remains difficult to compare analysis for experiments performed in different labs due to a lack of standardization.

Results

Here we propose a standardized single-molecule dataset (SMD) file format. SMD is designed to accommodate a wide variety of computer programming languages, single-molecule techniques, and analysis strategies. To facilitate adoption of this format we have made two existing data analysis packages that are used for single-molecule analysis compatible with this format.

Conclusion

Adoption of a common, standard data file format for sharing raw single-molecule data and analysis outcomes is a critical step for the emerging and powerful single-molecule field, which will benefit both sophisticated users and non-specialists by allowing standardized, transparent, and reproducible analysis practices.

Keywords

Single molecule, Standardized, File format, SMART, ebFRET, SMD

Background

Single-molecule techniques have proliferated over the past decade. Despite the power of these techniques and their widespread use, critical assessment of single-molecule data remains challenging. While there are multiple reasons for this, principal among these are the inherent noise and stochasticity associated with single-molecule events, which contribute substantially to the analysis challenge. To help manage similarly complex data sets generated from a number of techniques used in modern biological research, other fields have adopted standard data file formats, repositories, and analysis approaches. Examples include the PDB file format for structural data; the RCSB PDB repository of biomolecular structures; the NIH GenBank, DDBJ, and EMBL ENA repositories of gene and genome sequences; the NCBI BLAST and Ensembl sequence alignment and analysis tools; and the CNSsolve biomolecular

structure determination tool [15-24]. Standardization has been a key part of the development and advancement of these resources and techniques, facilitating data sharing and dissemination. In addition, the transparency of these formats, repositories, and tools encourages critical assessment of data. Individually the effect of these changes is difficult to assess, but cumulatively they contribute to increased reproducibility and reliability of measurements and, as a result, to the growth and widespread adoption of these techniques.

These examples represent important successes that have arisen naturally. However, several institutions and scientific leaders have recently begun to insist on greater transparency in the dissemination and treatment of all types of scientific data [25,26]. While there are many reasons for this desire and need, a number of well-documented instances within the drug discovery industry where the reproducibility of scientific results has been questioned [27-30] has raised awareness that a lack of easy access to raw data (arising from many sources) and a lack of tools for the primary analysis of the data can undermine clear communication of scientific results and can contribute to erroneous conclusions. Such high-profile problems cannot be attributed to any single failing, but a contributing cause is likely a current lack of standardization and control across the numerous measurement techniques that are combined to support these multidisciplinary development efforts [31,32].

Currently there is no standardization in place to unify the common aspects of most single-molecule data sets and to facilitate the use of the sophisticated analysis approaches that are continually being developed. We propose the single-molecule dataset (SMD) file structure as a general data format for storing and disseminating single-molecule data. Moreover, we take steps to facilitate this transition by making two previously established data-analysis packages created in independent labs compatible with this format.

Implementation

There are many commonalities in how single-molecule data are collected, stored, and analyzed. Figure 1A outlines three unifying relationships that form the basis of the SMD hierarchy. Most single-molecule datasets take the form of time series data (*i.e.*, traces) that are acquired simultaneously from one or more channels during an experiment. While this is not always the rawest form of the data (*e.g.*, a trace can be extracted from a movie recorded using a microscope that can simultaneously monitor many individual molecules), the single-molecule trace unifies many different techniques. At the highest level, a set of single-molecule traces (denoted as black rectangles in Figure 1A, top) are unified by the particular experiment that was used to generate them (denoted as a purple rectangle in Figure 1A, top). Finally, associated with each trace can be experimental information and quantities derived from the analysis of the raw single-molecule data (*e.g.*, inferred kinetic and thermodynamic parameters from model fitting; denoted as orange rectangle in Figure 1A, bottom). The aim of SMD is to encapsulate this hierarchy in a file structure that is independent of any particular programming language, data acquisition platform, or data analysis tool and that is widely compatible with distinct techniques and analysis strategies.

Figure 1 Structure of SMD. (A) Cartoon representation of the SMD hierarchy. (Top) Each experiment, represented by the purple rectangle, encompasses the raw data of many single-molecule traces, each represented by a black rectangle. (Bottom) Representation of an individual single-molecule trace within the above experiment. Raw single-molecule data consist of time series data arising from one or more channels. In this example, we depict two channels containing raw data as well as one channel containing an idealized trajectory

determined in post-processing. Associated with the raw data of each trace are attributes that are unique to that trace (depicted in orange), such as derived kinetic and thermodynamic parameters obtained from model fitting. **(B)** Representation of the SMD format in JavaScript Object Notation (JSON). The color scheme is used from the cartoon representation in panel (A).

There are many file types that easily accommodate the hierarchy of SMD (HDF5, .MAT, XML, etc.). Indeed, in any high-level analysis package one of these formats is likely to be used. However, to ensure the maximum interoperability between analysis tools, a standard text-based description is advantageous because it allows for straightforward determination of the data fields in a file without any prior knowledge of the specific experiment, data acquisition platform, or data analysis tools used. For interoperability purposes, a SMD object is represented in the widely used JavaScript Object Notation (JSON) format, whose nested structure naturally accommodates the SMD hierarchy.

Results and discussion

The SMD format aims to strike a balance between defining enough structure to facilitate interoperability of software packages and exchange of data between groups and providing enough flexibility to accommodate data associated with different experimental techniques and analysis use cases. The most important assumption we make is that the dataset holds traces with a fixed set of channels (*e.g.*, raw measurements, post-processed time series, inferred kinetic trajectories, etc.) that are annotated by some set of attributes (*e.g.*, pre-processing settings, fitted model parameters, etc.). The attributes may be quite specific to the type of experiment and analysis performed, but the channel values themselves should in general be suitable to visualization and analysis with different software packages. Figure 1B outlines how the three components of SMD are structured in the JSON notation (the top level is depicted in purple, raw data in black, and trace-specific parameters in orange). Each trace contains four fields. The *values* field stores the trace data where each data type is specified by a descriptive tag. The *index* field contains a list of row labels for the trace (typically measurement acquisition times). Any other trace-specific annotations (*e.g.*, pre-processing settings, fitted model parameters, etc.) are placed in the *attr* field. Finally the *id* field is used to store a 32 digit hexadecimal number generated by running the MD5 algorithm on the data for each trace. The list of traces is itself stored in the *data* field of an outer top-level structure, which itself has a dataset-specific *id* (generated by running the MD5 algorithm on the entire data structure) field as well as an *attr* field that holds top-level annotations or summary statistics that apply to the dataset as a whole (*e.g.*, experimental conditions, time and date of acquisition, averaged model parameters, etc.) and a *desc* field that contains a string describing the data set. Additionally, the dataset-specific *types* specifies the data type for each instance of data being stored in each set of *values.* A full description of the SMD specification is provided in the Supporting Material.

To facilitate the design and adoption of SMD we made the ebFRET [13,14] and SMART [11] single-molecule data analysis packages and visualization tools compatible with the SMD file format. We note here that ebFRET is a descendent of the previously released vbFRET [10,12] data analysis package. We also provide a number of tools for the basic support and validation of SMD files in both Matlab™ and Python packages. Full documentation of SMD and these tools is available at https://smdata.github.io/.

The collaboration that resulted in SMD enabled many details that are important for ensuring generality to be implemented. The ebFRET and SMART data analysis packages were developed independently from one another and as a result have significantly different functionalities and work flows. The ability of SMD to easily accommodate these packages with multiple graphical interfaces and distinct outputs provides a strong indication that SMD will be able to accommodate the needs of many researchers.

Conclusions

Adoption of SMD or, as needed, a different format that encapsulates generalities not anticipated at this time, is an important step for the realization of the full potential of single-molecule measurements by and for a broad scientific community. Although it will require some discipline for researchers to abide by (or "follow") a common set of standards, the potential long-term benefits are hard to overstate. Standardization will help facilitate the transfer of information among different labs by ensuring that a minimal structure and set of information are present. In turn, this information sharing will facilitate further critical assessment (*e.g.*, data quality, error assessment, and reproducibility) and reanalysis of single-molecule datasets, important steps in extracting the most from complex but information-rich single-molecule data. Moreover, adoption of a common data standard could help facilitate the creation of a repository for single-molecule data (analogous to the RCSB PDB repository of biomolecular structures), which would enable a high degree of transparency and would ensure that data obtained now yields further insights in years to come. We are hopeful that the flexibility of SMD can easily accommodate the needs of current researchers and that it will enable researchers to reap the benefits that accompany widely adopted standardization.

Availability and requirements

Project name: Single-molecule dataset (SMD)

Project home page: https://smdata.github.io/

Operating system: Platform independent

Programing Languages: Support provided for Matlab™ and Python, but SMD is not tied to any particular programing language.

Other requirements: none

Licenses: creative commons

Any restrictions to use by non-academics: none

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MG, JMW, DSP, HM, CHW, RLG and DH all contributed to the inception of the project. MG, JWM and DSP carried out the design and implementation of the SMD format. MG, JMW, DSP, HM, CHW, RLG and DH all contributed to the writing of the manuscript. MG updated the SMART package to be compatible with SMD and JWM updated ebFRET to be compatible with SMD. All authors read and approved the final manuscript.

Acknowledgments

The authors would like to thank any members of the single-molecule community who take the time to adopt the SMD format. In particular we would like to thank Prof. Frederick Sacks for agreeing to make the widely used QuB analysis package compatible with the SMD format and for Prof. Taekjip Ha for agreeing to make the widely used HaMMy analysis package compatible with the SMD format. Additionally we would like to thank members of the Herschlag and Gonzalez labs as well as Prof. Aaron Hoskins (University of Wisconsin at Madison) for critical feedback. This work was supported by a NIH PPG Grant (GM49243) to D.H.; a NSF CAREER Award (MCB 0644262) and a NIH National Institute of General Medical Sciences grant (R01 GM084288) to R.L.G.; a NIH National Centers for Biomedical Computing grant (U54CA121852) to C.H.W.; a Rubicon fellowship (680-50-1016) from the Netherlands Organization for Scientific Research (NWO) to J.W.M.; and a NIH training grant in Biotechnology (5T32GM008412) to M.G.

References

1. Joo C, Fareh M, Kim VN: **Bringing single-molecule spectroscopy to macromolecular protein complexes.** *Trends Biochem Sci* 2013, **38:**30–37.

2. Dulin D, Lipfert J, Moolman MC, Dekker NH: **Studying genomic processes at the single-molecule level: introducing the tools and applications.** *Nat Rev Genet* 2013, **14:**9–22.

3. Coltharp C, Yang X, Xiao J: **Quantitative analysis of single-molecule superresolution images.** *Curr Opin Struct Biol* 2014, **28C:**112–121.

4. Woodside MT, Block SM: **Reconstructing folding energy landscapes by single-molecule force spectroscopy.** *Annu Rev Biophys* 2014, **43:**19–39.

5. Liu Y, Park J, Dahmen KA, Chemla YR, Ha T: **A comparative study of multivariate and univariate hidden Markov modelings in time-binned single-molecule FRET data analysis.** *J Phys Chem B* 2010, **114:**5386–5403.

6. Qin F, Auerbach A, Sachs F: **A direct optimization approach to hidden Markov modeling for single channel kinetics.** *Biophys J* 2000, **79:**1915–1927.

7. McKinney SA, Joo C, Ha T: **Analysis of single-molecule FRET trajectories using hidden Markov modeling.** *Biophys J* 2006, **91:**1941–1951.

8. Qin F, Auerbach A, Sachs F: **Hidden Markov modeling for single channel kinetics with filtering and correlated noise.** *Biophys J* 2000, **79:**1928–1944.

9. Watkins LP, Yang H: **Information bounds and optimal analysis of dynamic single molecule measurements.** *Biophys J* 2004, **86:**4015–4029.

10. Bronson JE, Fei J, Hofman JM, Gonzalez RL Jr, Wiggins CH: **Learning rates and states from biophysical time series: a Bayesian approach to model selection and single-molecule FRET data.** *Biophys J* 2009, **97:**3196–3205.

11. Greenfeld M, Pavlichin DS, Mabuchi H, Herschlag D: **Single Molecule Analysis Research Tool (SMART): an integrated approach for analyzing single molecule data.** *PLoS One* 2012, **7:**e30024.

12. Bronson JE, Hofman JM, Fei J, Gonzalez RL Jr, Wiggins CH: **Graphical models for inferring single molecule dynamics.** *BMC Bioinformatics* 2010, **11**(8)**:**S2.

13. Van de Meent J-W, Bronson JE, Wiggins CH, Gonzalez RL Jr: **Empirical Bayes methods enable advanced population-level analyses of single-molecule FRET experiments.** *Biophys J* 2014, **106:**1327–1337.

14. Van de Meent J-W, Bronson JE, Wood F, Gonzalez RL Jr, Wiggins CH: **Hierarchically-coupled hidden Markov models for learning kinetic rates from single-molecule data.** *Proc Int Conf Mach Learn* 2013, **28:**361–369.

15. McGinnis S, Madden TL: **BLAST: at the core of a powerful and diverse set of sequence analysis tools.** *Nucleic Acids Res* 2004, **32**(Web Server issue)**:**W20–W25.

16. Brünger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL: **Crystallography & NMR system: A new software suite for macromolecular structure determination.** *Acta Crystallogr D Biol Crystallogr* 1998, **54**(Pt 5)**:**905–921.

17. Dolinski K, Ball CA, Chervitz SA, Dwight SS, Harris MA, Roberts S, Roe T, Cherry JM, Botstein D: **Expanding yeast knowledge online.** *Yeast Chichester Engl* 1998, **14:**1453–1469.

18. Sherlock G, Hernandez-Boussard T, Kasarskis A, Binkley G, Matese JC, Dwight SS, Kaloper M, Weng S, Jin H, Ball CA, Eisen MB, Spellman PT, Brown PO, Botstein D, Cherry JM: **The Stanford Microarray Database.** *Nucleic Acids Res* 2001, **29:**152–155.

19. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28:**235–242.

20. Berman HM: **The Protein Data Bank: a historical perspective.** *Acta Crystallogr A* 2008, **64**(Pt 1)**:**88–95.

21. Tateno Y, Imanishi T, Miyazaki S, Fukami-Kobayashi K, Saitou N, Sugawara H, Gojobori T: **DNA Data Bank of Japan (DDBJ) for genome scale research in life science.** *Nucleic Acids Res* 2002, **30:**27–30.

22. Hamm GH, Cameron GN: **The EMBL data library.** *Nucleic Acids Res* 1986, **14:**5–9.

23. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank.** *Nucleic Acids Res* 2013, **41:**D36–D42.

24. Bilofsky HS, Burks C: **The GenBank genetic sequence data bank.** *Nucleic Acids Res* 1988, **16**(5 Pt A)**:**1861–1863.

25. Tibshirani R: **Big data: how to avoid a big mess**. .

26. **Reducing our irreproducibility**. *Nature* 2013, **496**:398.

27. Tibshirani R: **Immune signatures in follicular lymphoma.** *N Engl J Med* 2005, **352:**1496–1497. author reply 1496–1497.

28. Ioannidis JPA: **Why most published research findings are false.** *PLoS Med* 2005, **2:**e124.

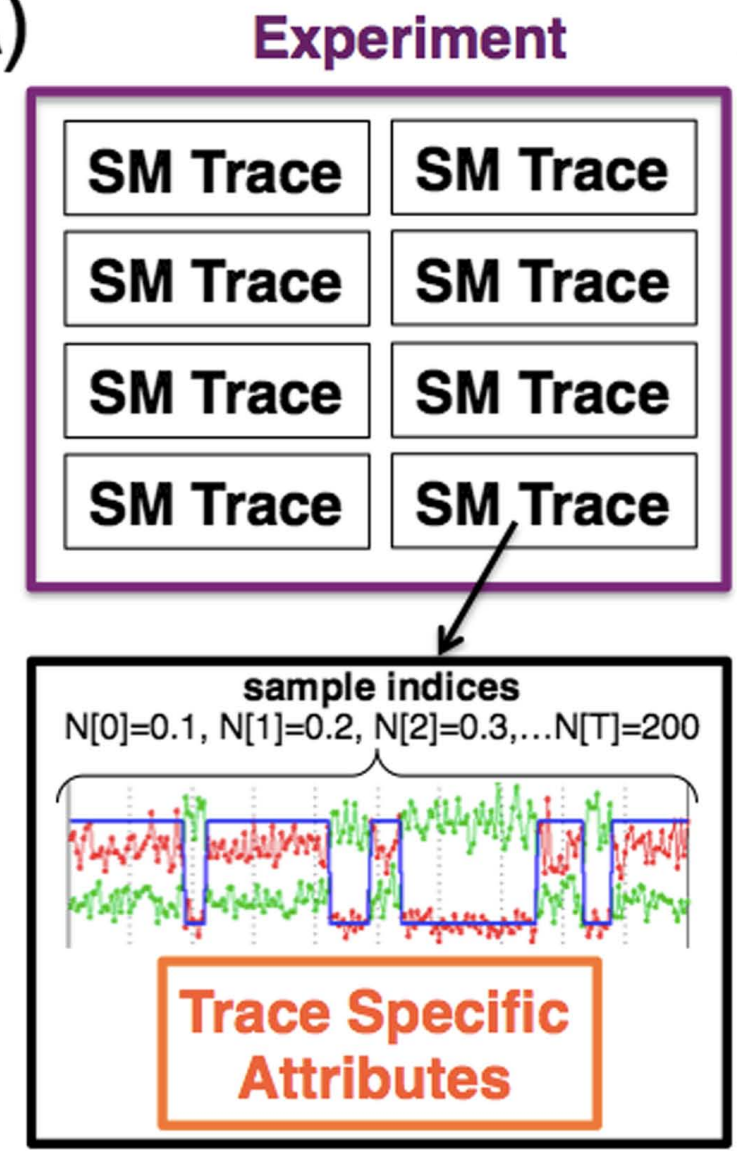
29. Begley CG, Ellis LM: **Drug development: Raise standards for preclinical cancer research.** *Nature* 2012, **483:**531–533.

30. Prinz F, Schlange T, Asadullah K: **Believe it or not: how much can we rely on published data on potential drug targets?** *Nat Rev Drug Discov* 2011, **10:**712.

31. Ioannidis JPA: **How to make more published research true.** *PLoS Med* 2014, **11:**e1001747.

32. Ioannidis JPA, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, Moher D, Schulz KF, Tibshirani R: **Increasing value and reducing waste in research design, conduct, and analysis.** *Lancet* 2014, **383:**166–175.

A)



B)

```
{
    "desc": "dataset descriptor",
    "id": "dataset hash",
    "attr": {
        "data_package": "ACME Analysis"},
    "types": {
        "index": "format specifier",
        "values": {"col1": "format specifier",
                   "col2": "format specifier",
                   "col3": "format specifier",...}},
    "data": [
        {
            "id": "32 character random hash tag",
            "index": [N0,N1,N2,...NT],
            "values": [
                "col1": [X0, X1, X2,... XT],
                "col2": [Y0, Y1, Y2,... YT],
                "col3": [Z0, Z1, Z2,... ZT],...],
            "attr": {
                "Variable name 1": String or Number,
                "Variable name 2": String or Number,...}
        },
        {
            "id": ...
            "index": ...
            "values": ...
            "attr": ...
        },
      ...
    ]
}
```

Additional files provided with this submission:

Additional file 1. (208k)
http://www.biomedcentral.com/content/supplementary/s12859-014-0429-4-s1.docx