Computer Physics Communications ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Computer Physics Communications

journal homepage: www.elsevier.com/locate/cpc

Fast-NPS—A Markov Chain Monte Carlo-based analysis tool to obtain structural information from single-molecule FRET measurements☆

Tobias Eilert [a], Maximilian Beckers [b], Florian Drechsler [a], Jens Michaelis [a,*]

[a] *University of Ulm, Institute for Biophysics, Albert-Einstein-Allee 11, 89081 Ulm, Germany*
[b] *European Molecular Biology Laboratory, Structural and Computational Biology Unit, Meyerhofstrasse 1, 69117 Heidelberg, Germany*

ARTICLE INFO

Article history:
Received 6 April 2017
Received in revised form 24 May 2017
Accepted 28 May 2017
Available online xxxx

Keywords:
Bayesian inference
Structural biology
smFRET
Nano-Positioning system
Dye model

ABSTRACT

The analysis tool and software package *Fast-NPS* can be used to analyse smFRET data to obtain quantitative structural information about macromolecules in their natural environment. In the algorithm a Bayesian model gives rise to a multivariate probability distribution describing the uncertainty of the structure determination. Since *Fast-NPS* aims to be an easy-to-use general-purpose analysis tool for a large variety of smFRET networks, we established an MCMC based sampling engine that approximates the target distribution and requires no parameter specification by the user at all. For an efficient local exploration we automatically adapt the multivariate proposal kernel according to the shape of the target distribution. In order to handle multimodality, the sampler is equipped with a parallel tempering scheme that is fully adaptive with respect to temperature spacing and number of chains. Since the molecular surrounding of a dye molecule affects its spatial mobility and thus the smFRET efficiency, we introduce dye models which can be selected for every dye molecule individually. These models allow the user to represent the smFRET network in great detail leading to an increased localisation precision. Finally, a tool to validate the chosen model combination is provided.

Programme summary

Programme Title: *Fast-NPS*

Programme Files doi: http://dx.doi.org/10.17632/7ztzj63r68.1

Licencing provisions: Apache-2.0

Programming language: GUI in MATLAB (The MathWorks) and the core sampling engine in C++

Nature of problem: Sampling of highly diverse multivariate probability distributions in order to solve for macromolecular structures from smFRET data.

Solution method: MCMC algorithm with fully adaptive proposal kernel and parallel tempering scheme.

1. Introduction

Structural biology is a key field of life sciences. It aims to give a picture of life at the molecular level, which is a prerequisite for obtaining a mechanistic molecular understanding of cellular processes. Modern techniques, such as *X-ray* crystallography or *cryo-EM*, can resolve the structure of macromolecules up to an Ångström level. However, in order to understand nature it is important to not only resolve the static structure of macromolecules in artificial environments, but also to elucidate their dynamic structure or even transient dynamic complexes in an aqueous milieu. Here, an interesting tool is *Förster resonance energy transfer* (*FRET*) [1]. FRET has been termed a *molecular ruler*, since the distance range that can be measured is on the length scale of proteins [2]. Furthermore, *single-molecule FRET* (*smFRET*) has become a widely used technique for studying the dynamics of macromolecular complexes [3–5]. While distance information can be obtained, determining distances quantitatively remains a challenge [6–8]. Using the trilateration of several smFRET distances one can determine an unknown position with respect to a macromolecular complex [9,10]. Moreover, smFRET results have been combined with other techniques for structural biology approaches [11–13].

We present the *Fast-Nano-Positioning System* (*Fast-NPS*), an advanced software package that utilises smFRET measurements between *dye molecules* (further referred to simply as *dyes*) to gain quantitative structural information about macromolecular

☆ This paper and its associated computer programme are available via the Computer Physics Communication homepage on ScienceDirect (http://www.sciencedirect.com/science/journal/00104655).

* Corresponding author.
E-mail address: jens.michaelis@uni-ulm.de (J. Michaelis).

http://dx.doi.org/10.1016/j.cpc.2017.05.027
0010-4655/

complexes. It aims to localise the unknown dye positions, the so-called *antennas*, by means of dyes covalently bound to known positions on a macromolecular complex, the *satellites*, with a Bayesian model. In Bayesian statistics the degree of knowledge about an unknown quantity is expressed by a probability distribution conditional on the evidence obtained from experimental data, the so-called *posterior* [14].

In *Fast-NPS* the data constitutes the measured average smFRET efficiencies with experimental error, the determined isotropic Förster radii and the measured steady-state fluorescence anisotropies [15,16]. The unknown quantities of interest are the positions of the antennas. Since not only the positions, but also the orientations of the dyes have a tremendous effect on the measured smFRET efficiencies, their *transition dipole moments* (*TDMs*) are explicitly modelled in *Fast-NPS*.

In previous works the *Nano-Positioning System* (*NPS*) has already been used to study the position of the exiting RNA from the eukaryotic RNA polymerase II [10] and to investigate the influence of the transcription factor TFIIB on the position of the nascent RNA [11]. Further, the position of the non-template and upstream DNA in yeast Polymerase II transcription elongation complexes [17] and the architecture of a minimal Polymerase II open promoter complex [18] were analysed. Moreover, *NPS* was also applied to shed light on the archaeal initiation complex [19].

In Section 2 we recapitulate the Bayesian framework of *NPS* which forms the basis for the analysis of smFRET data in order to gain structural information from macromolecular complexes. In Section 3.1 we present a novel *Markov Chain Monte Carlo* (*MCMC*) algorithm, which forms the basis for *Fast-NPS*, to analyse the posterior in order to extract structural information from an arbitrarily large smFRET network using standard desktop computers in a reasonable amount of time. In Section 3.2 we establish the theory of individual dye models, accounting for the different spatial behaviour over time depending on the dye's molecular environment. A method to assess the consistency of the prior information with the smFRET data is presented in Section 3.3. In Section 4 we apply the dye models to a real smFRET network guided by the help of our consistency check in order to maximise the localisation precision. In Section 5 we give a comprehensive discussion of Sections 3–4 and end with a conclusion and a future outlook in Section 6.

2. Theory

In Bayesian statistics the normalised posterior $\mathcal{P}$ is proportional to the product of *likelihood* $\mathcal{L}$ and *prior* Π, i.e. $\mathcal{P} \propto \mathcal{L} \cdot \Pi$. The likelihood is the probability distribution of the parameters, in our case the dye positions and orientations, given the experimental data and a statistical model. According to the central limit theorem the distribution of the average smFRET efficiency $\langle E \rangle$ converges to a normal distribution centred around $\mathbb{E}(E)$, the expectation of the smFRET efficiency E, when the number of data points becomes large. Assuming that there is a unique configuration of dyes giving rise to the data, $\mathbb{E}(E)$ is the associated smFRET efficiency. Thus, the likelihood function is defined by a normal distribution centred around the average smFRET efficiency with standard deviation σ, where σ is the experimentally determined measurement error, i.e. we have $\mathcal{L} := \mathcal{N}(\langle E \rangle, \sigma^2)$[1] [15]. Further, the dependence of the average smFRET efficiency $\langle E \rangle$ on the distance d between donor and acceptor and their *Förster distance* R is given by

$$\langle E \rangle = \frac{1}{1 + (d/R)^6}. \tag{1}$$

[1] For simplicity both random variables and their realisations will be denoted with the same symbols throughout this paper.

The Förster distance R is given by

$$R = R^{\text{iso}} \sqrt[6]{\frac{3}{2}\kappa^2}, \tag{2}$$

where the *orientation factor* κ^2 is a function of the positions and TDMs of the dye molecules [1]. Depending on the relative orientation of donor and acceptor it can take values from 0–4 [20]. When both dye molecules are free to rotate and reorient faster than the fluorescence lifetime of the donor in the presence of the acceptor, κ^2 adopts its isotropic value of 2/3, such that all orientation effects vanish. Under this condition the Förster distance is called the *isotropic Förster distance* R^{iso} [21]. The position of a dye i is parametrised by x_i, y_i, z_i and its TDM by an azimuthal angle θ_i and a polar angle φ_i. Substitution of Eq. (1) into the likelihood represents a transformation from the range of the average smFRET efficiency to the *configuration space* Ω of both dyes. The parameter vector of the likelihood defined on Ω is then $\boldsymbol{x} := (x_1, y_1, z_1, \theta_1, \varphi_1, x_2, y_2, z_2, \theta_2, \varphi_2)^T$.

Fast-NPS is a hybrid approach. The prior knowledge about the position of the dye molecule is gained from an *accessible volume* (*AV*) computation with respect to the structure of the molecule of interest obtained from *X-ray* crystallography, *NMR* studies or *cryo-EM* [11]. These volumes serve as a flat position prior for *Fast-NPS*. Finally, the prior for the TDMs is uniform over the unit hemisphere (see Section 3.1.3).

Since the unambiguous localisation of one or more antennas requires the measurement against several satellites, the joint likelihood is given by an uncorrelated (assuming independent measurements) multivariate normal distribution in the space of the average smFRET efficiencies. The joint prior consists then of the product of the uniform distributions on the individual AVs. Finally, according to Bayes' law the joint posterior is proportional to the product of joint likelihood and joint prior [15]. The parameter vector is then given by $\boldsymbol{x} := (x_1, y_1, z_1, \theta_1, \varphi_1, \ldots, x_n, y_n, z_n, \theta_n, \varphi_n)^T$, where n denotes the number of dyes in the network.

In the following section we focus on the development of an adaptive sampling engine for the structural inference of a large variety of smFRET networks.

3. Methods

3.1. Algorithm

In Section 2 we have defined a probability distribution on the configuration space Ω providing us with the information how likely a realisation $\boldsymbol{x}$ is. The structural information, which we can extract from the posterior, is given by a volume which specifies how likely it is that a certain dye position is found inside. Although the posterior can be written in a closed form, the marginalisation down to the position of one dye is analytically unfeasible. In order to solve this problem we chose to develop a sampling algorithm which produces a set of realisations $\{\boldsymbol{x}\}$, so-called *samples*, drawn from the posterior. Then, the marginalisation is reduced to the simple projection onto the positions of the dye of interest.

Since we want to express a complete lack of position knowledge about an antenna when starting the analysis, the search space covers usually more than 25 times the volume of the macromolecule. However, the major posterior mass which is displayed by a credible volume, i.e. the smallest volume that includes a certain probability, covers only a small fraction of this search space (Fig. 1A). In order to efficiently draw samples from the posterior, we have chosen an *importance sampling algorithm*, or more specifically, a *Metropolis-within-Gibbs* sampler [22].

Localisation geometry enforced by limitations of biochemical labelling strategies can also induce banana- or spherical shell-like

Please cite this article in press as: T. Eilert, et al., *Fast-NPS*—A Markov Chain Monte Carlo-based analysis tool to obtain structural information from single-molecule FRET measurements, Computer Physics Communications (2017), http://dx.doi.org/10.1016/j.cpc.2017.05.027.