Available online at www.sciencedirect.com
SCIENCE DIRECT®

Computer Physics Communications 158 (2004) 150–157

Computer Physics Communications

www.elsevier.com/locate/cpc

FRETsg: Biomolecular structure model building from multiple FRET experiments ☆

G.F. Schröder, H. Grubmüller *

Theoretical and Computational Biophysics Group, Max-Planck-Institute for Biophysical Chemistry, 37077 Göttingen, Germany

Received 4 July 2003; accepted 4 February 2004

Abstract

Fluorescence energy transfer (FRET) experiments of site-specifically labelled proteins allow one to determine distances between residues at the single molecule level, which provide information on the three-dimensional structural dynamics of the biomolecule. To systematically extract this information from the experimental data, we describe a program that generates an ensemble of configurations of residues in space that agree with the experimental distances between these positions. Furthermore, a fluctuation analysis allows to determine the structural accuracy from the experimental error.

Program summary

Title of program: FRETsg
Catalogue identifier: ADTU
Program obtainable from: CPC Program Library, Queen's University of Belfast, N. Ireland
Program summary URL: http://cpc.cs.qub.ac.uk/summaries/ADTU
Computer: SGI Octane, Pentium II/III, Athlon MP, DEC Alpha
Operating system: Unix, Linux, Windows98/NT/XP
Programming language used: ANSI C
No. of bits in a word: 32 or 64
No. of processors used: 1
No. of bytes in distributed program, including test data, etc.: 11407
No. of lines in distributed program, including test data, etc.: 1647
Distribution format: gzipped tar file
Nature of the physical problem: Given an arbitrary number of distance distributions between an arbitrary number of points in three-dimensional space, find all configurations (set of coordinates) that obey the given distances.
Method of solution: Each distance is described by a harmonic potential. Starting from random initial configurations, their total energy is minimized by steepest descent. Fluctuations of positions are chosen to generate distance distribution widths that best fit the given values.

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

* Corresponding author.
E-mail address: hgrubmu@gwdg.de (H. Grubmüller).
URL: http://www.mpibpc.gwdg.de/abteilungen/070.

0010-4655/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.cpc.2004.02.001

© 2004 Elsevier B.V. All rights reserved.

PACS: 87.15.By; 33.15.Bh; 33.50.-j; 87.64.Ni

Keywords: Fluorescence resonance energy transfer; FRET; Protein structure; Structural fluctuations; Structure determination; Distance constraints; Molecular conformation

1. Introduction

Fluorescence Resonance Energy Transfer (FRET) is a powerful tool to investigate the structure of biomolecules, such as DNA or proteins. FRET experiments yield the distance between two fluorescent dyes, a donor and an acceptor, which can be covalently attached at specific positions to the biomolecule and thus provide information on the structure of that biomolecule [1–3]. Recent advances in both, spectroscopy and molecular biology, allow to carry out multiple FRET experiments, that yield a whole set of distances between different positions in the biomolecule [4]. In a typical experiment to investigate the conformation of a protein, double cysteine mutants of the protein, produced by site directed mutagenesis, are labelled with cysteine binding donor and acceptor dyes, e.g., Alexa488 and Alexa594, respectively. This method has been applied, e.g., to the protein Syntaxin 1A, where 15 site directed double mutants have been produced. These allowed to measure 15 intramolecular distances, which were used to build a model of the protein structure [5].

Single molecule FRET experiments additionally contain information on distance distributions. The widths of these distributions originate from the flexibilities of the dyes, from the conformational flexibility of the biomolecule, and from shot noise. To transform these distances into three-dimensional models of the biomolecule, we implemented an approach similar in spirit to the one widely applied in structure determination from NMR experiments [6]. However, several generalizations were necessary, because single molecule FRET experiments yield distances over a much larger range than NMR and additionally give information on distance distributions. Thus, a strict fluctuation analysis is provided to obtain also a more detailed description of the conformational flexibility of the biomolecule and of the dye. Furthermore, the Förster radius, which is a crucial parameter in the calculation of the distance from the measured donor and acceptor intensities, often cannot be determined exactly. Therefore, the dependence of the resulting model on the chosen Förster radius can be studied.

2. Theory

2.1. Fluorescence resonance energy transfer

If a fluorescent dye, called a donor, is excited, energy is transferred to a nearby acceptor dye by the Förster mechanism [7]. This energy transfer depends on the overlap of the emission spectrum of the donor and the absorption spectrum of the acceptor as well as on the distance between the dyes and their relative orientation. Accordingly, by measuring donor and acceptor fluorescence intensities, I_D and I_A, the distance r between the two dyes is obtained, usually via

$$\frac{I_A}{I_A + I_D} = \frac{1}{1 + (r/r_0)^6}, \tag{1}$$

where r_0 is the dye-specific effective Förster radius, which typically also includes (averaged) dye orientation effects [8].

2.2. Outline of method

Here, each distance is assigned a harmonic potential term, resulting in a potential function V,

$$V(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \sum_{\text{FRET pairs } i,j} \left(|\mathbf{x}_i - \mathbf{x}_j| - d_{ij}\right)^2, \tag{2}$$

where $\mathbf{x}_i$ denotes the position of the dye at residue i, and d_{ij} is the measured distance between position i and j, determined from (1). Subsequently, the potential V is minimized by varying the positions $\mathbf{x}_i$ using gradient descent.

Depending on the number and values of the provided distances, the optimum is usually not unique. To sample these possible multiple solutions, the minimization is repeated many times from random initial configurations. Thereby, a sufficiently large ensemble of configurations obeying the given distances is obtained.

If some positions are already known, e.g., from X-ray experiments, the distances between them can be used to further improve the model. With FRETsg, this is achieved by setting "known" flags of the respective positions in the positions file (∗.pos). The distances between all "known" positions are then calculated and override those distances given in the FRET distances file (∗.dis) (see below). Note that after conversion into distances, these optional coordinates in the positions file are discarded and not used for the further calculations.

Optionally, donor and acceptor *intensities* can be given together with a Förster radius R_0 in which case the respective distance is calculated using (1). This provides a convenient way to study the dependency of the model on the chosen Förster radius.

2.2.1. Representation of the ensemble of configurations

The above described algorithm yields an ensemble of configurations. A common way to represent such an ensemble is to fit the individual configurations onto each other. To optimize the resulting representation and depending on the kind of structural information the user is interested in, it can be useful to select only a subset of atoms to be used for the fitting procedure. This is achieved by "fit" flags.

2.2.2. Handedness

Since only distances are available, for each solution structure also its mirror image is a valid solution. To avoid the resulting mixing of configurations with differing handedness, the following procedure has been implemented. First, for each obtained structure, the rmsd (root mean square deviation) of both the original and the mirrored structure from the first configuration, which defines the reference handedness, is calculated. Then, from each pair of configurations, the one with the lower rmsd value is selected and added to the set of resulting configurations.

2.3. Fluctuation analysis

Within the usual approximation of isotropic fluctuations of the residue positions, the standard deviations of the distances are transformed into standard deviations of position fluctuations according to

$$\sigma_{D,ij}^2 = \frac{1}{2\pi\sigma_i\sigma_j} \int |\mathbf{x}_i - \mathbf{x}_j|^2 \, \mathrm{e}^{-|\mathbf{x}_i - \mathbf{x}_i^0|^2/2\sigma_i^2} \, \mathrm{e}^{-|\mathbf{x}_j - \mathbf{x}_j^0|^2/2\sigma_j^2} \, \mathrm{d}\mathbf{x}_i \, \mathrm{d}\mathbf{x}_j = \sigma_i^2 + \sigma_j^2, \tag{3}$$

where $\sigma_{D,ij}^2$ is the measured variance of the distance between position i and j, and σ_i^2 and σ_j^2 are the variances (which are to be determined) of the fluctuations of positions i and j, respectively, and $\mathbf{x}_i^0$ denotes the configuration obtained by minimizing V as defined in (2). Note that the obtained variances are independent of the $\mathbf{x}_i^0$.

To calculate the position fluctuations σ_k of each position k, the following scoring function S,

$$S = \sum_{\text{FRET pairs } i,j} \left[(\sigma_i^2 + \sigma_j^2) - \sigma_{D,ij}^2\right]^2, \tag{4}$$

is minimized using gradient descent.

Note that the widths of the intensity distributions originate from both, uncorrelated fluctuations like shot noise or detector noise, and from the conformational flexibility of the dyes and the biomolecule, which leads to correlated intensity fluctuations. Therefore, to infer the width of a distance distribution from the widths of the intensity distributions, the correlation between the donor and the acceptor intensity would have to be considered. Because such correlation analysis is outside the scope of this program, and to avoid over-interpretation of the data, no fluctuation analysis is performed, if any of the distances is input via intensities.

2.4. Input data

FRETsg reads three input files:

2.4.1. The parameter file

The parameter file (∗.inp) contains filenames of further input and output files as well as all parameters used in the program.

```
#
# This is a sample input file for
# FRETsg v1.0
#
sample.pos     Positions
sample.dis     Distances
sample.pdb     Output PDB-File
50             Number of Configurations to create (max. 2000)
0.005          Gradient Threshold of Minimization (0.0 - 1.0)
TRUE           Fluctuation Analysis [TRUE/FALSE]
TRUE           Write Mirror Structure [TRUE/FALSE]
FALSE          Write POV-Ray output [TRUE/FALSE]
```

2.4.2. The positions file

In the positions file (∗.pos) all position labels must be defined. Positions that should be used to fit the resulting configurations to each other are marked here by setting the “fit” flag to “1”. The distances between all positions with given coordinates (in Å) and “known” flag set are used in the model building. The respective distances in the distances file are then ignored.

```
# This is a sample positions file
# for FRETsg v1.0
#
#   known   fit   X     Y     Z
#Label            [ only if known=1 ]
#
10    1      1   3.0   4.0   5.0
20    1      1   5.0   1.0  10.0
30    0      1  -3.4   5.4  21.0
40    0      1
50    0      0
```

Note that for label 30 the xyz coordinates are ignored.

2.4.3. The distances file

The distances file (∗.dis) contains the given distances (in Å) and their standard deviations.

```
# This is a sample Distances File
# for FRETsg v1.0
#
#
#   Mode = D : using Distance
#   Mode = I : using Intensities
#
#
#   Id : Donor Intensity
#   Ia : Acceptor Intensity
#   RO : Forster Radius (in Ang)
#
#   Label2 Mode Distance Width   Id   Ia    RO
#Label1
10  20      I     30.0    2.3   178  1000  40.0
10  30      D     30.0    2.2  1445   654  47.0
20  30      D     30.0    1.4
10  40      D     25.0    2.2
20  40      D     25.0    1.2
30  40      D     25.0    0.7
10  50      D     40.0    2.0
20  50      I     40.0    1.1   312  1190  50.0
30  50      D     40.0    0.9
```

If Mode = “I”, then the distance is calculated via (1) using the given I_D, I_A and R_0 values. This overrides the value given in the “Distance”-column. Note, if any of the distances are defined by the intensities, no fluctuation analysis is performed.

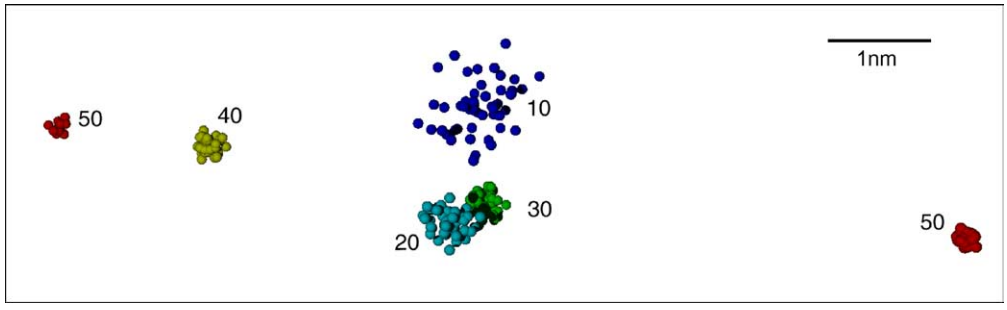

Fig. 1. Output of FRETsg (*sample.pdb*) using the sample input files (see text). The coloured clusters represent the determined positions. The cluster labels correspond to the example positions file (*sample.pos*). Note that position 50 falls into two clusters, as explained in the text. The sizes of the clusters are determined by the fluctuation analysis (see text).

2.5. Output data

The main output is in PDB-format [9], which is the common format to represent protein structures. The output-file contains all calculated configurations. Thus, the three-dimensional structure can be viewed with any available protein structure viewer, e.g., Rasmol [10]. The potential energy (or "strain") of each position appearing in (2) after the minimization is given in columns 10 and 11 (designated for occupancy and temperature factor), such that high values, which show inconsistencies in the set of distances, can be highlighted by color. Fig. 1 shows an example obtained from the given input files shown above.

Optionally, the mirror images of all configurations can be written out. These represent also a solution, which possibly has to be taken into account.

For the preparation of high quality images, an input file for the raytracer *POV-Ray*^{tm} [11] is optionally produced.

Acknowledgements

We thank Claus Seidel and Enno Schweinberger for helpful discussions. Support by the Volkswagen-Foundation is gratefully acknowledged.

Appendix A. Test run

The input files for the test run are *sample.inp*, *sample.pos* and *sample.dis*. In this example, five labels and nine distances are defined. Apparently, with the given distances, two different clusters fulfill the experimental constraints. Depending on which atom positions are used for fitting the structures onto each other, this ambiguity will show up at different positions. In the given example, position 50 is not used for the fit, and therefore falls into two clusters (see Fig. 1).

The following output is produced by FRETsg:

```
$ fretsg sample.inp

     FRETsg 1.0
     (C) 2003 Gunnar F. Schroder
     3D Model Builder from multiple FRET distances.

Reading input parameters from file sample.inp ...
Reading position labels from file sample.pos ...
Reading distance pairs from file sample.dis ...

Creating 50 minimized Configurations .... Done.
Run Fluctuation Analysis ................ Done.

standard deviations of fluctuating positions :
---------------------------------------------
Position 10 : 2.048
Position 20 : 1.080
Position 30 : 0.747
Position 40 : 0.572
Position 50 : 0.382
```

```
Writing Output to file sample.pdb.
Writing Mirror Output to file sample.mirror.pdb.

$
```

This test run should output two files; *sample.pdb* and *sample_mirror.pdb.*

```
sample.pdb:

MODEL     1
ATOM    1 C GLN 10    1.187 -14.382  6.479 0.00 0.00   A
ATOM    2 C GLN 20  -14.922   1.427 -9.318 0.00 0.00   A
ATOM    3 C GLN 30   14.889   5.028 -8.718 0.00 0.00   A
ATOM    4 C GLN 40   -1.826   7.126  8.472 0.00 0.00   A
ATOM    5 C GLN 50   -3.822  22.166 22.112 0.00 0.00   A
ENDMDL
MODEL     2
ATOM    1 C GLN 10    3.622 -18.229  7.539 0.00 0.00   A
ATOM    2 C GLN 20  -15.310   2.107 -9.489 0.00 0.00   A
ATOM    3 C GLN 30   15.022   5.419 -8.009 0.00 0.00   A
...

sample_mirror.pdb:

MODEL     1
ATOM    1 C GLN 10   -1.187 -14.382  6.479 0.00 0.00   A
ATOM    2 C GLN 20   14.922   1.427 -9.318 0.00 0.00   A
ATOM    3 C GLN 30  -14.889   5.028 -8.718 0.00 0.00   A
ATOM    4 C GLN 40    1.826   7.126  8.472 0.00 0.00   A
```

```
ATOM    5 C GLN 50    3.822  22.166 22.112 0.00 0.00   A

ENDMDL

MODEL     2

ATOM    1 C GLN 10   -3.622 -18.229  7.539 0.00 0.00   A

ATOM    2 C GLN 20   15.310   2.107 -9.489 0.00 0.00   A

ATOM    3 C GLN 30  -15.022   5.419 -8.009 0.00 0.00   A

...
```

References

[1] T. Ha, Methods 25 (2001) 78.
[2] D.D. Root, X. Shangguan, J. Xu, M.A. McAllister, J. Struct. Biol. 127 (1999) 22.
[3] T. Yasunaga, Y. Suzuki, R. Ohkura, K. Sutoh, T. Wakabayashi, J. Struct. Biol. 132 (2000) 6.
[4] S. Weiss, Science 283 (1999) 1676.
[5] M. Margittai, J. Widengren, E. Schweinberger, G. Schröder, S. Felekyan, E. Haustein, M. König, D. Fasshauer, H. Grubmüller, R. Jahn, C.A.M. Seidel, Proc. Natl. Acad. Sci. USA 100 (2003) 15516.
[6] M. Nilges, Curr. Op. Struct. Biol. 6 (1996) 617.
[7] T. Förster, Ann. Phys. 2 (1948) 55.
[8] B.W. van der Meer, G. Cooker, S.-Y. Chen, Resonance Energy Transfer: Theory and Data, VCH Publishers, New York, 1994.
[9] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The protein data bank, Nucl. Ac. Res. 28 (2000) 235–242.
[10] R. Sayle, E.J. Milner-White, Trends in Biochemical Sciences 20 (9) (1995) 374.
[11] Persistence of Vision^{tm} Ray Tracer, electronic access: http://www.povray.org.