Supplementary Information for

Unraveling the complexity of the interactions of DNA nucleotides with gold by single molecule force spectroscopy

Fouzia Bano, Damien Sluysmans, Arnaud Wislez, Anne-Sophie Duwez[1]

Data analysis

The majority of the collected data sets show a distinct tail to the right, i.e. the force histograms are skewed towards high force. This is a common feature of rupture force measurements and it has been explained by invoking multiple attachment events[1] or heterogeneity in the chemical bonding and dynamical disorder[2, 3]. Most probably, all these features contribute to the observed statistics of the rupture force. In order to account for multiple binding and other anomalies, we fit the recorded data with a Gaussian mixture model (GMM), that is, a weighted sum of M component Gaussian densities as given by the equation

$$P(\mathbf{x}|\lambda) = \sum_{i=1}^{M} p_i G(\mathbf{x}|\mu_i, \sigma_i) \qquad (1)$$

where x is the vector of the observed rupture forces at a given loading rate, $G(\mathbf{x}|\mu_i, \sigma_i)$ is the normalized Gaussian component with mean μ_i and variance σ_i and p_i is the weight of the i-th component. The weights satisfy the normalization condition $\sum_{i=1}^{M} p_i = 1$, and λ represents the set of all the parameters $\lambda = \{p_i, \mu_i, \sigma_i\}$ for i=1,…M that specify the model. The total probability density in eq.(1) allows the treatment of cases in which the observed data in x are sampled from

[1] To whom correspondence should be addressed

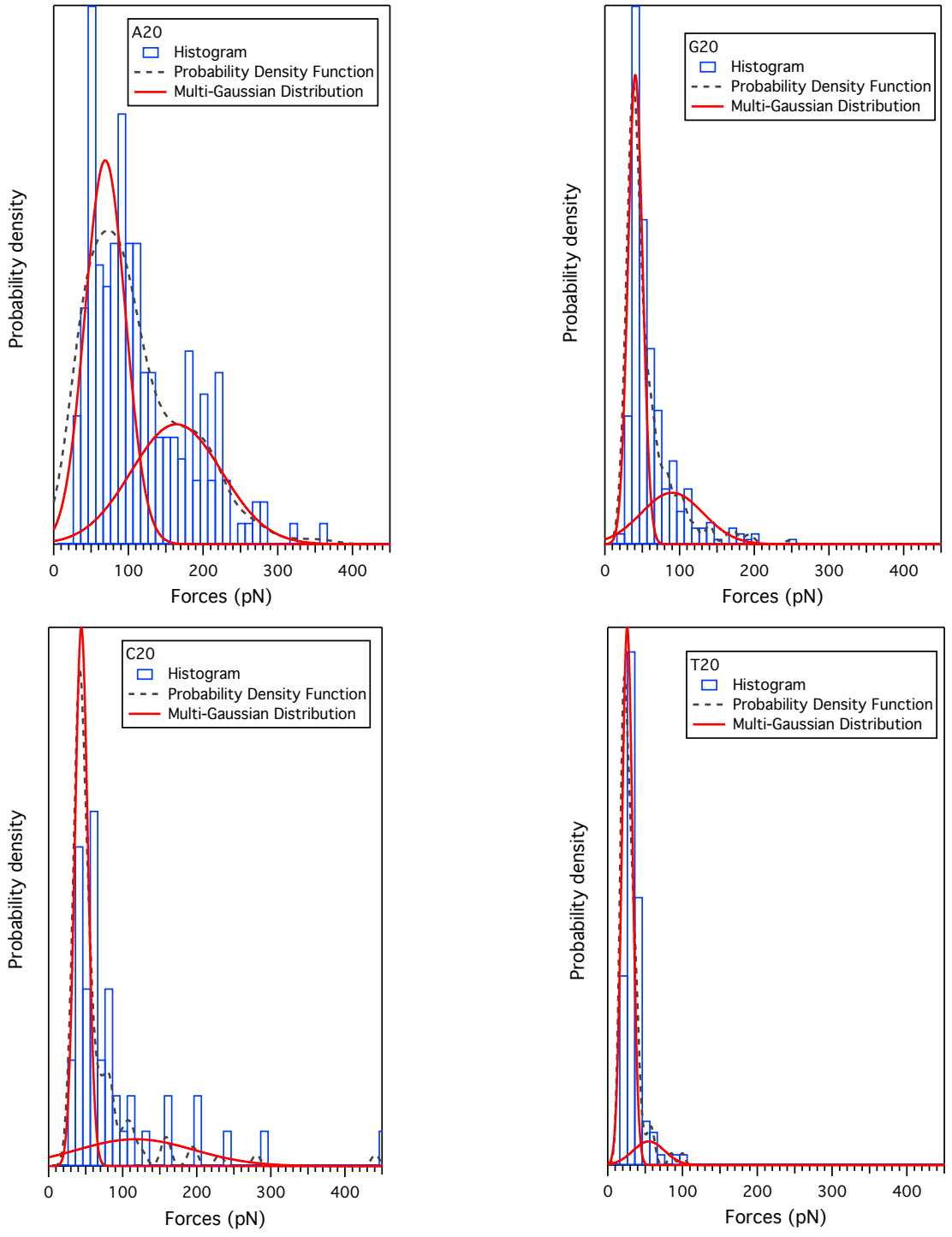
M sub-populations, but no information is available about the sub-population to which an individual observation belongs. The mixture model is used to make statistical inferences about the properties of the sub-populations given only observations on the pooled population, without sub-population identity information. The set of parameters $\lambda = \{p_i, \mu_i, \sigma_i\}$ for i=1,…M are determined through the Expectation-Maximization algorithm[4] that maximize the likelihood of the GMM given the observed data through an iterative procedure. The GMM provides a smooth overall distribution fit of the observed data and its components can be used to detail the multi-modal nature of the density.

In practice, we used a two component Gaussian model to account for the populations observed in the probability density function graphs of all the experiments. We used the Gaussian component with the higher weight to determine the most probable rupture force, while we identify the other components as describing the population of outliers composed of multiple bond rupture events (when we observe a double or triple force) or to another binding mode.

The figures below show the obtained two-mode Gaussian density superimposed to the histograms of all the rupture force data sets and the Probability Density Function (PDF). For each estimated average rupture force, the ± 95% confidence interval was computed as $\pm 2\sqrt{\sigma^2/(p_1 N)}$ where σ^2 is the estimated variance of the main Gaussian component while $p_1 N$ represents the effective size of the population.

For A_{20}, the main component accounts for 59% of the overall observed population. The second Gaussian accounts for the rest. Two maxima clearly appear from the PDF (dotted line). For G_{20}, C_{20} and T_{20}, the PDF shows (besides the main peak which accounts for 70-90% of the overall population) a series of several maxima of very small intensity. This series is brought together in

the second Gaussian. Those maxima correspond to multiples of the main peak and can be interpreted as the multiple attachment events[1] or heterogeneity in the chemical bonding and dynamical disorder[2, 3] mentioned above.



Supporting Figures

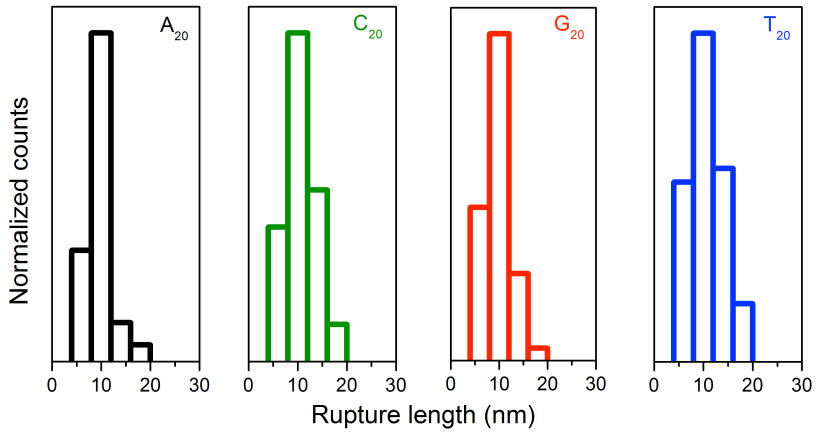


Figure S1: Histograms of the rupture lengths for A20 (black), C20 (green), G20 (red) and T20 (blue).

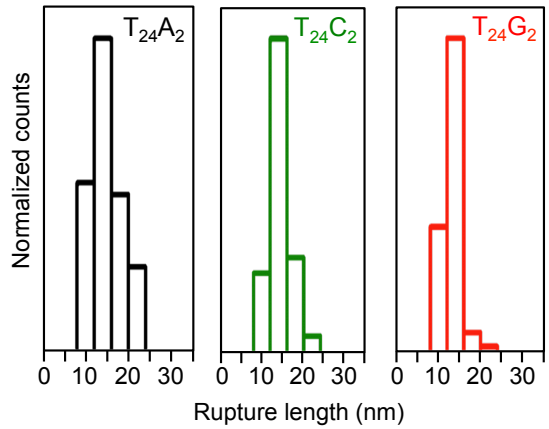


Figure S2: Histograms of the rupture lengths for T24A2 (black), T24C2 (green), T24G2 (red)

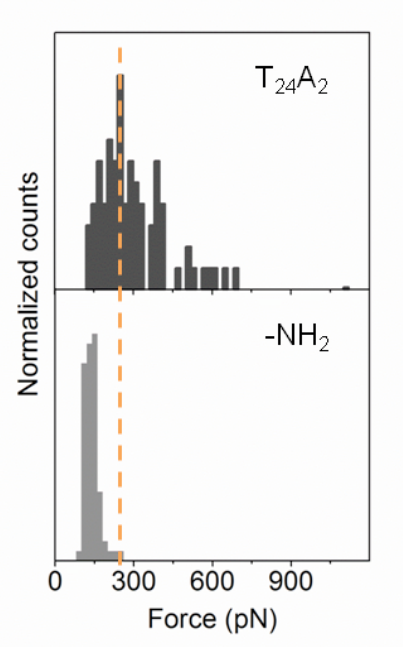


Figure S3: Control experiment. Histograms of the rupture forces for adenine-Au (upper panel) and amine-Au (lower panel). Experiments were done in pure water, 0.01s dwell time. The dashed line is provided as a guide to the eyes. The amine group was coupled to the AFM tip through a 22bp heterogeneous ssDNA spacer molecule. The results show that the interactions between adenine and Au are complex and not only due to the free amine group of adenine.

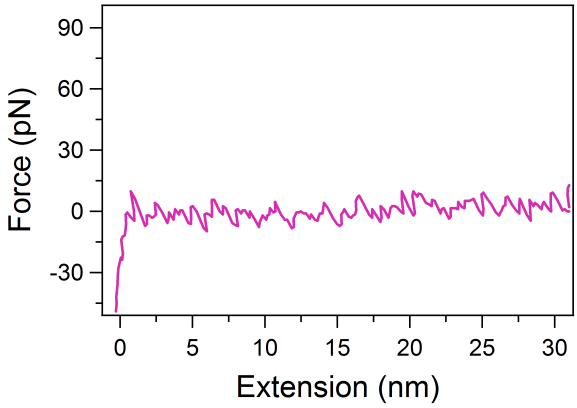


Figure S4: Control experiment. Force extension curve (retraction), registered at 100 nm/s between a tip functionalized with MCH and gold in 150 mM NH_4OAc aqueous solution. No specific peaks are observed.

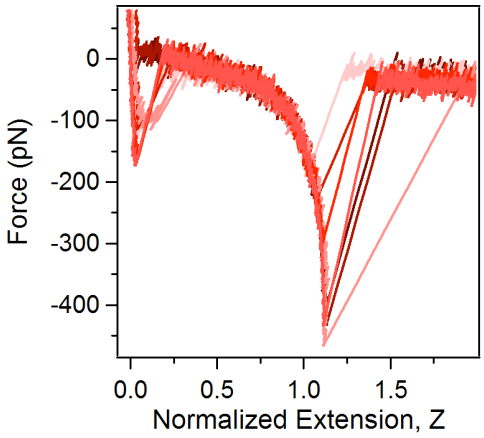


Figure S5: Force-extension curves of T24G2 in water at 1s dwell time, normalized and superimposed to rule out the possibility of the stretching and rupturing of multiple base-Au bonds. 6 T24G2 curves were normalized by setting the extension (Z) to 1nm at F = 150 pN

Calculation of DNA oligomer length

The theoretical length of the oligomers was determined using the base length or distance of 0.6 nm (ref. 5) between two bases and MCH linker of 0.65 nm (ref. 6 and references therein). Using these values, the theoretical lengths of $T_{24}X_2$ and X_{20} are 16.3 nm and 12.6 nm, respectively.

References

1. S. Guo, C. Ray, A. Kirkpatrick, N. Lad and B. B. Akhremitchev, *Biophys J*, 2008, **95**, 3964-3976.
2. M. Raible, M. Evstigneev, F. W. Bartels, R. Eckel, M. Nguyen-Duong, R. Merkel, R. Ros, D. Anselmetti and P. Reimann, *Biophys J*, 2006, **90**, 3851-3864.
3. C. Hyeon, M. Hinczewski and D. Thirumalai, *Phys Rev Lett*, 2014, **112**, 138101.
4. A. P. Dempster, N. M. Laird and D. B. Rubin, *Journal of the Royal Statistical Society. Series B (Methodological)*, 1977, **39**, 1-38.
5. W. S. Chen, W. H. Chen, Z. Chen, A. A. Gooding, K. J. Lin and C. H. Kiang, *Phys Rev Lett*, 2010, **105**, 218104.
6. D. Scaini, M. Castronovo, L. Casalis and G. Scoles, *ACS Nano*, 2008, **2**, 507-515.