A Generalized Mathematical Model To Estimate T- and B-Cell Receptor Diversities Using AmpliCot

Irina Baltcheva,[†] Ellen Veel,[‡] Thomas Volman,[‡] Dan Koning,[‡] Anja Brouwer,[‡] Jean-Yves Le Boudec,[†] Kiki Tesselaar,[‡△] Rob J. de Boer,[§] and José A. M. Borghans[‡]

[†]Laboratory for Computer Communications and Applications, École Polytechnique Fédérale de Lausanne, Switzerland; and [‡]Department of Immunology, University Medical Center Utrecht, and [§]Department of Theoretical Biology, Utrecht University, Utrecht, The Netherlands

Supplementary Material

S1. Fit of Time-Series Data

The best-fits of both the second order kinetics and the heteroduplex model to the time-series annealing kinetics of data sets 1-3 are given in Figure S1, Figure S2 and Figure S3.

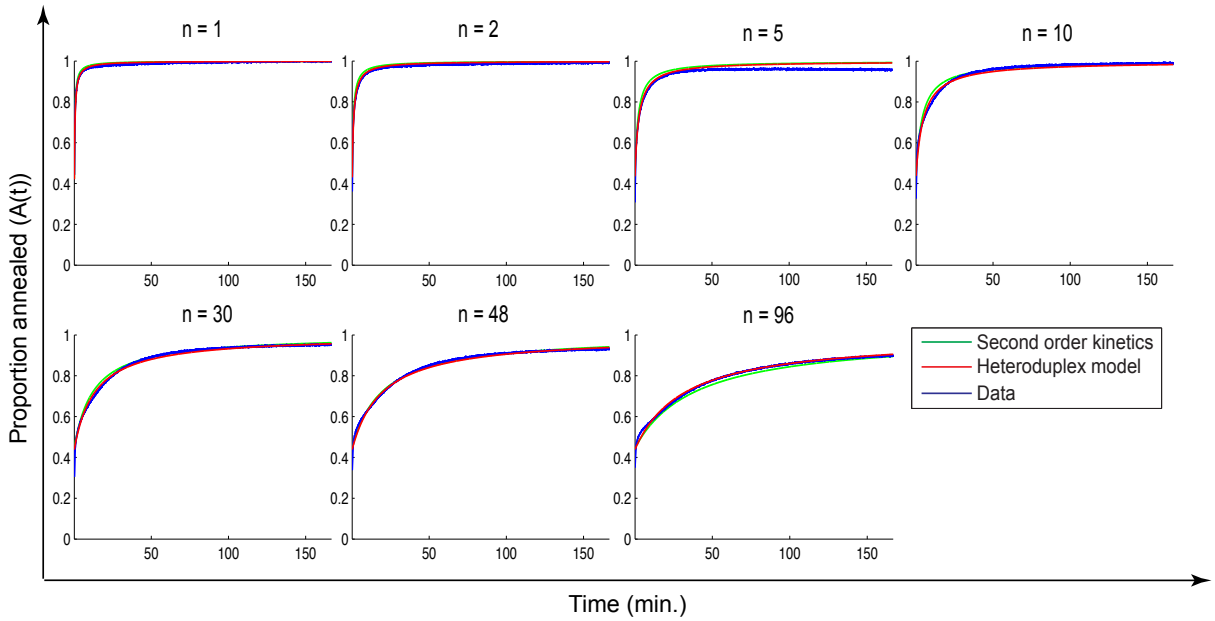

Figure S1: Best-fit of Baum&McCune's data [1, Fig.2a] for known diversity templates (each panel). Blue: experimental data. Green: best-fit of the second order kinetics model. Red: best-fit of the heteroduplex model. For the best-fit parameters and confidence intervals, see Table S1. The heteroduplex model fits significantly better than second order kinetics (p-value $< 10^{-3}$).

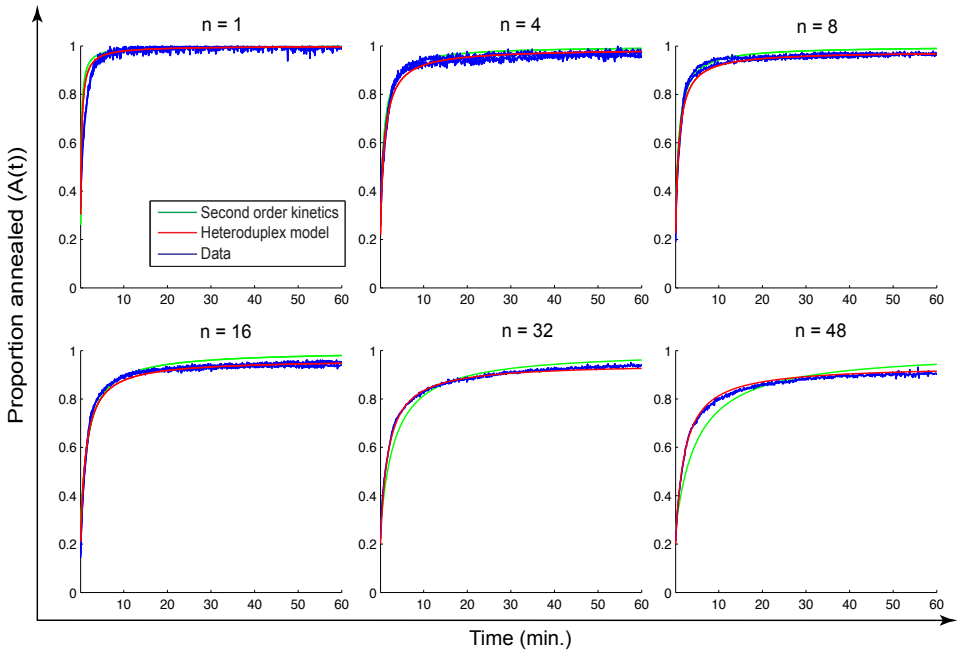

Figure S2: Best-fit of data set 2 for known diversity templates (each panel). Blue: data sample (two replicates). Green: best-fit of second order kinetics. Red: best-fit of the heteroduplex model. For the best-fit parameters and confidence intervals, see Table S1. The heteroduplex model fits significantly better than second order kinetics (p-value $< 10^{-3}$).

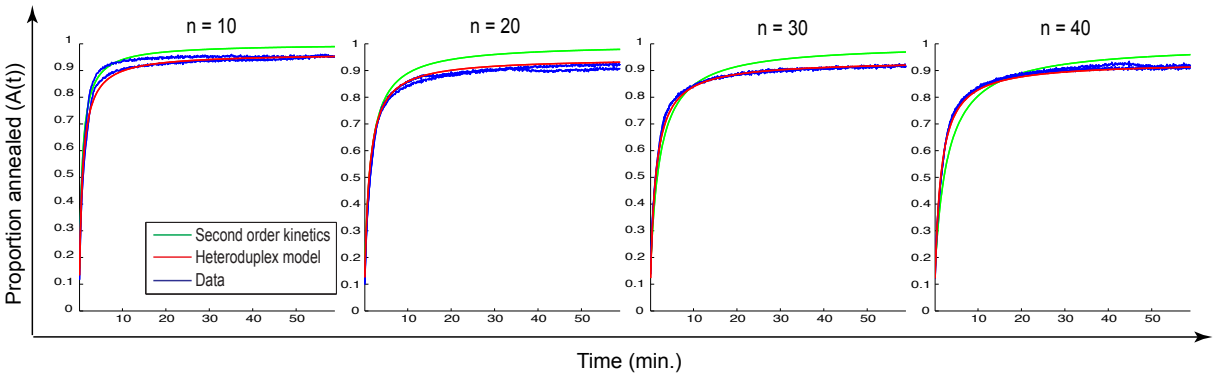

Figure S3: Best-fit of data set 3 for known diversity templates (each panel). Blue: data sample (two replicates). Green: best-fit of second order kinetics. Red: best-fit of the heteroduplex model. For the best-fit parameters and confidence intervals, see Table S1. The heteroduplex model fits significantly better than second order kinetics (p-value $< 10^{-3}$).

S2. Best-Fit Parameters: Time-Series Data

The best-fit parameters of both models fitted to the different time-series data of Figure S1, Figure S2 and Figure S3) are given in Table S1.

Model	Param.	Data set 1		2		3	
		Value	95% CI	Value	95% CI	Value	95% CI
SOK	ML	35 577		10 835		7 051	
	a	2.2160	[2.1835, 2.2535]	12.9887	[12.3851, 13.6816]	7.9439	[7.6179, 8.2326]
	α	0.5532	[0.5462, 0.5605]	0.7435	[0.7123, 0.7741]	0.8423	[0.8210, 0.8588]
HM	ML	39 444		13 635		11 343	
	a	2.1719	[1.7439, 3.7591]	9.3091	[9.0372, 16.4313]	6.5503	[3.9625, 9.5867]
	α	0.5608	[0.5524, 0.5697]	0.8180	[0.7795, 0.8497]	0.8930	[0.8715, 0.9443]
	ξ_1	0.7962	[0.4639, 1.0000]	0.9994	[0.5195, 1.0000]	0.6275	[0.3813, 0.9982]
	ξ_2	0.0093	[0.0052, 0.0112]	0.0923	[0.0521, 0.0946]	0.0884	[0.0534, 0.1397]
	φ	0.9703	[0.9482, 0.9946]	0.8915	[0.8824, 0.8970]	0.8969	[0.8912, 0.9051]

Table S1: Best-fit reaction rates and 95% confidence intervals (CI). Each model was fitted to the annealing data of data sets 1-3 (Figure S1, Figure S2, Figure S3) by minimizing the sum of squared errors (on log scale). ML is the maximum likelihood of the best-fit. $\xi_1 = \frac{z_1}{(z_1+d_1)}$ and $\xi_2 = \frac{z_2}{(z_2+d_2)}$. The confidence intervals (in parentheses next to the parameter values) were computed using 999 bootstrap replicates [2]. SOK: second order kinetics, HM: heteroduplex model.

For all data sets the association rate a was lower under the heteroduplex model than under second order kinetics, whereas the proportion α of melted molecules was higher under the heteroduplex model than under second order kinetics. Since the composite parameter $\xi_1 = \frac{z_1}{(z_1+d_1)}$ was above 0.5, i.e., the zipping rate z_1 was larger than the dissociation rate d_1, transient homoduplexes had a higher probability to permanently hybridize than to dissociate. Because $\xi_2 = \frac{z_2}{(z_2+d_2)} < 0.5$, transient heteroduplexes tended to dissociate rather than to hybridize completely.

Of interest are the quantitative differences between the best-fit parameters of the three data sets. First, the proportion of melted molecules α was larger for data sets 2 and 3 than for data set 1. Possibly, more time elapsed between the effective start of the annealing phase and the first measurement in data set 1. Second, we observed that ξ_2 in data set 1 is one order of magnitude smaller than in the other data sets. This means that in data set 1, only a few heteroduplexes were formed. The closer ξ_2 is to 0, the larger d_2 (the dissociation of transient heteroduplexes) is compared to z_2 (the zipping of heteroduplexes). Finally, the fluorescence of heteroduplexes in the first data set is only slightly lower than the fluorescence of homoduplexes ($\varphi \approx 0.97$), whereas in data sets 2 and 3, it is about 90% of that of homoduplexes ($\varphi \approx 0.9$). This difference could be caused by the structural difference of the oligonucleotides used in the three data sets.

S3. Best-Fit Parameters: Cot Data

Table S2 compares the best-fit parameters of both models fitted to annealing data (AD, as in Table S1), or to Cot 50% or 80% values (as in Figure 6). Note that we did not fit Cot values of data set 3 because of the low number of different diversities in this data set. Similarly, we did not fit annealing curves of data set 4 because this data set was used as a validation set to our diversity prediction procedure. The fitted parameters are rather different according to the method used (annealing data vs Cot values). This was expected in the case of second order kinetics because parameters a and α are not identifiable in Eq. (11). The discrepancies observed between the fits of Cot 50% and Cot 80% values highlight the importance of the choice of an annealing percent. The data (and consequently the best-fit parameters) exhibited very different properties according to which point was chosen (Figure 6). Fitting the annealing curves (time-series data) has the advantage of being independent of the choice of an annealing percentage.

Model	Par.	Data set								
		1			2			3	4	
		AD	Cot 50%	Cot 80%	AD	Cot 50%	Cot 80%	AD	Cot 50%	Cot 70%
SOK	a	2.2160	6.3647	2.2549	12.9887	3.8716	13.1807	7.9439	7.0568	45.0009
	α	0.5302	0.5748	0.7276	0.7435	0.5542	0.4998	0.8423	0.5056	0.5002
HM	a	2.1719	5.1001	2.9702	9.3091	4.6209	13.1994	6.5523	380.9901	38.4029
	α	0.5608	0.7122	0.5644	0.8180	0.6100	0.5074	0.8930	0.9283	0.5935
	ξ_1	0.7962	0.7121	0.5957	0.9994	0.7440	0.8045	0.6275	0.9998	0.9901
	ξ_2	0.0093	0.0381	0.0301	0.0923	0.0513	0.0120	0.0884	0.0001	0.0001
	φ	0.9703	0.6664	0.6790	0.8915	0.7653	0.7918	0.8969	0.2860	0.6060

Table S2: Comparison of the best-fit parameters as fitted on time-series annealing data (AD, as in Table S1), or directly on Cot 50% or 80% values (Figure 6 of main text). Each model was fitted to the data by minimizing the sum of squared errors between Cot values and Eq. (10) (HM: heteroduplex model) or Eq. (11) (SOK: second order kinetics).

References

[1] P. D. Baum and J. M. McCune. Direct measurement of T-cell receptor repertoire diversity with AmpliCot. *Nature Methods*, 3:895 – 901, Oct 2006.

[2] J. Carpenter and J. Bithell. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine*, 19(9):1141–1164, 2000.