Folding Stabilizes the Evolution of Catalysts

Stephan Altmeyer
Rudolf M. Füchslin
John S. McCaskill
Fraunhofer Gesellschaft
Biomolecular Information Processing
Institute Center Birlinghoven
D-53754 St. Augustin, Germany
mccaskill@biomip.fhg.de

Abstract Sequence folding is known to determine the spatial structure and catalytic function of proteins and nucleic acids. We show here that folding also plays a key role in enhancing the evolutionary stability of the intermolecular recognition necessary for the prevalent mode of catalytic action in replication, namely, *in trans*, one molecule catalyzing the replication of another copy, rather than itself. This points to a novel aspect of why molecular life is structured as it is, in the context of life as it could be: folding allows limited, structurally localized recognition to be strongly sensitive to global sequence changes, facilitating the evolution of cooperative interactions. RNA secondary structure folding, for example is shown to be able to stabilize the evolution of prolonged functional sequences, using only a part of this length extension for intermolecular recognition, beyond the limits of the (cooperative) error threshold. Such folding could facilitate the evolution of polymerases in spatially heterogeneous systems. This facilitation is, in fact, vital because physical limitations prevent complete sequence-dependent discrimination for any significant-size biopolymer substrate. The influence of partial sequence recognition between biopolymer catalysts and complex substrates is investigated within a stochastic, spatially resolved evolutionary model of trans catalysis. We use an analytically tractable nonlinear master equation formulation called PRESS (McCaskill et al., *Biol. Chem.* 382: 1343–1363), which makes use of an extrapolation of the spatial dynamics down from infinite dimensional space, and compare the results with Monte Carlo simulations.

Keywords
Evolution, catalysis, RNA, altruism, origins of life, recognition

1 Introduction

The information dynamics of self-replicating systems has gained steadily in significance since the pioneering work of Manfred Eigen [7]. The issue here is not the historical question of the origin of the first self-replicating system, but the scientific question of whether, following a chance inception of such a system, molecular kinetics can support the evolution of information necessary for complex catalysis. Even if a simple self-replicating system is established by chance, it is not evident that such a system will be stable against the pressure of mutations and decay. The quasispecies model [7, 9] addressed this issue for simple self-replicating molecules with the non-interacting kinetics of the form $X \rightarrow 2X$. However, it is still unclear whether the chemistry necessary for such simple (cis-acting) self-replication (for instance with an RNA-like molecule) could be sustained at sufficient rates to compete with destruction events in the prebiotic world. The evolution of multimolecular replication systems with efficient and

specialized sequence-encoded catalysts, however, harbors a second major problem, that of exploitation, which is the main issue of this article.

Catalysts that act not on themselves (i.e., reflexively on the same molecular copy) but on other substrates are the rule in biochemistry (examples include polymerase enzymes and ligases), and we call them *trans-acting*. Trans-acting catalysts may still be autocatalytic, in that their substrate is another copy of the same molecular type, but the normal situation is that catalysts act on a different type of substrate molecule. Early work on coupled catalytic sets of molecules addressed the issues of kinetic stability, sequence encoding, and evolvability under idealized assumptions of non-exploited sequence dependences in catalytic hypercycles [9] and incrementally evolvable catalytic sets [15]. Today it is generally accepted that spatial resolution enables local stabilization of interacting and replicating molecules against the destructive force of evolving exploitation [4, 5]. Here, exploitation can arise by mutation when the mutated offspring of a replicator, which can act as a polymerase catalyst for replication, are no longer catalytically active but continue to be replicated in the system by existing replicator polymerases. If the system reaches a state in which the production of non-catalysts exceeds that of the catalysts locally, exploitation critically depletes catalytic activity and for finite resources the system is doomed to collapse locally.

With spatial isolation and discrete individuals, simpler modes of coupled catalysis than the hypercycle have been shown to achieve evolutionary stability [6, 19, 20]. In particular, an analytically tractable stochastic model of the evolution of catalysis has been developed [21] exhibiting exploitation control in the simplest trans-acting autocatalytic model $2X \rightarrow 3X$. The catalyst exploiting species Y is replicated by X according to $X + Y \rightarrow X + 2Y$ and arises also by mutation. It was shown that even catalysts without specific self-recognition (i.e., unable to discriminate between their own templates and those of the exploiting species) can be sustained in the course of evolution. This is a second level of neutral evolution, one involving generic catalysts, and expected to be the most common type of problem that early molecular systems had to deal with. In fact, even exploiting species Y, which are replicated faster by X than X itself, can be kept at bay under suitable spatial conditions, namely, for appropriate molecular diffusion coefficients. Summarizing these facts, the $2X \rightarrow 3X$ reaction scheme is now referred to as the *generous replicator* model.

In the current article, the focus is on the way in which molecular recognition between a catalyst and its target can help to stabilize the spatial evolution of trans-acting catalysis. We shall compare the effects of three degrees of folding: no folding, *partial* folding, and *complete* folding. Without folding, local molecular recognition is perfectly correlated with sequence, whereas complete folding is defined as a random mapping in that the recognition elements of the folded structures are arranged in a manner that is uncorrelated with sequence similarities. Correlation functions for more realistic folding, exemplified by algorithmic prediction of RNA secondary structure, have been shown [8, 13] to have an intermediate degree of correlation between sequence and structure, and can be deemed partial folding.

There are parallels here to the widespread study of kin recognition as a common factor in the evolution of altruism in population biology (see [11] and [1] for a review in the molecular context). However, clearly defined physical mechanisms for molecular recognition can add precision to the analysis of the sequence dependence of recognition and its role in evolutionary biology. In particular, the molecular context provides explicit physically grounded correlation limits to the sequence dependences of kin recognition. The situation is more generally of interest in biology because of the rather generic cooperative replication mechanism, $2X \rightarrow 3X$, which we shall study in the presence of mutant variants Y that, depending on the sequence, may or may not be discriminated by X (i.e., whether or not $X + Y \rightarrow X + 2Y$ depends on molecular

structural details of the sequences X and Y). We shall investigate the way in which molecular folding mediates the sequence dependence of molecular recognition and its influence on the evolutionary stability of catalysis. We shall utilize existing large scale studies of sequence dependent RNA folding [13] to confirm the quantitative validity of our general conclusions in a more specific context. We shall show that the basic stability induced by spatial resolution is enhanced considerably when sequence folding occurs, even for very limited local recognition of molecules.

For evolutionary stability, complete sequence recognition of relevant templates for replication would be optimal. However, besides being unphysical, such a system would suffer from its lack of both generality and evolvability. Limited recognition, on the other hand, has to deal with the problem of deleterious small sequence changes going undetected and accumulating as replicated junk in the system. If the recognition is at the level of folded structures, the situation may be greatly improved, since the effect of a single mutation can be distributed over the entire structure (altering the minimum-free-energy structure, for example), and may be detected by a recognition mechanism with limited information as input. In this sense, folding acts similarly to check sums in information technology, convolving information to allow local testing.

To analyze these effects in a tractable evolutionary model, the PRESS platform (probability reduced evolution of spatially resolved species [2, 21]) is used, where an infinite number of sites are arranged in such a way that every site is directly connected with every other one (simplex topology). On each site, a discrete and bounded number of molecules are situated with reactions only taking place locally, but single molecules can migrate between any pair of sites. The PRESS model provides a simple abstraction of spatial structure, as in Wright's island model of evolution [24], but its consequences may be mapped onto continuous physical spaces in various dimensions. The topology is common to the model of Kimura [16], but details differ: the PRESS framework is solvable explicitly for small site bounds. A compact description of PRESS is given for reference in Appendix 2. Later on, Monte Carlo simulations are performed to verify the PRESS results as well as to obtain numerical results for lower dimensions and for the case of a completely resolved sequence space.

For the self-contained case of polymerases acting trans-catalytically on templates, it will be shown that the sensitivity of folded structures to global sequence changes, though incomplete, allows an evolutionary effective discrimination of encoding template sequences, in contrast to unfolded recognition schemes. The situation for proteins is expected to be similar. The net effect is to give catalysts an additional stability against exploiting mutants, allowing an enlargement of the sequence length and opening the way to the evolution of more complex catalysts. It would be beneficial to incorporate such features in more abstract models of evolving life.

2 Methods

2.1 Evolution of Generalist Polymerases without Recognition

The elementary reactions of the generous replicator model [21], which occur at each site in a spatially distributed population (see Appendix 2), are

$$\begin{aligned} \mathrm{X}+\mathrm{X}+\emptyset &\xrightarrow{Q} \mathrm{X}+\mathrm{X}+\mathrm{X},\\ \mathrm{X}+\mathrm{X}+\emptyset &\xrightarrow{1-Q} \mathrm{X}+\mathrm{X}+\mathrm{Y},\\ \mathrm{X}+\mathrm{Y}+\emptyset &\xrightarrow{1} \mathrm{X}+\mathrm{Y}+\mathrm{Y},\\ \mathrm{X},\,\mathrm{Y} &\xrightarrow{d} \emptyset \end{aligned} \tag{1}$$

Here, trans-acting catalytic replication of all species by the catalyst X is normalized to occur at a unit rate (by a simple rescaling of time). In this most demanding scenario, only one sequence, X, has catalytic function. All others are taken to be equally non-functional and can therefore be commonly denoted by Y. As with simple self-replication, the quality factor $Q < 1$ gives the probability of error-free replication. The processes of loss with a rate coefficient d and molecular interchange with a rate D (see also Appendix 2) complete the reaction kinetics of the model. Vacant locations (symbol ∅) are required for reactions in this kinetic model, summarizing the influence of saturation effects (resource limitations or product inhibition) that limit the catalytic reactions on each site.

With non-vanishing mutation rates, the non-structured system goes extinct due to the synthetic advantage of the exploiters Y. However, the system can be stabilized against mutation in spatially resolved environments (even in the ∞-dimensional simplex case) for appropriate diffusion rates, $D_{\infty,\min} \leq D_{\infty} \leq D_{\infty,\max}$ [22]. This means that despite the fact that non-functional sequences Y have a selective advantage on a local site (by a factor Q^{-1} in the neutral case), the functional catalysts persist in the population. For neutral or hyper-exploitation, however, the stabilization against mutation was shown to be relatively weak (with $Q_{\text{crit}} > 0.9$). While chemical recognition is incomplete in nature, we shall see that even a limited amount of local molecular recognition, if structured appropriately in sequence space, can provide a major enhancement of the information capacity (number of coding bits of information preserved in the system) of an evolving catalytic system. We retain the most demanding scenario, where only a single functionally active sequence occurs, to highlight the effects of recognition and reduce the complexity of the analysis. Owing to the way in which recognition changes the viability domain of the generous replicator model, it turns out that these results can easily be projected onto the recognition system.

2.2 Influence of Folding on Template Recognition

In order for two biomolecules to react, they must first dock to one another, reflecting a *fit* or *induced fit* in some space. This lock and key procedure physically involves only a fraction of the whole sequence: the recognition domain. Full recognition, where every monomer of the target sequence is distinguished, would lead to an unphysical 100% distinction between sequences. Sequence recognition is incomplete in nearly every biomolecular reaction, with the possible exception of double helical nucleic acid hybridization for short sequences. In the case of RNA catalysts, such as ribozyme ligases, target recognition occurs predominantly via Watson-Crick reverse complementary base pairing between portions of the catalyst and portions of its target. Recognition between folded structures can be reduced again to a simple local string matching procedure, but now between two strings representing the folded structures. Three cases are investigated: no folding, complete folding, and an intermediate case based on RNA secondary structure. In each case, the initial population involves a well-adapted catalyst X able to copy the template molecule that encodes it (X), but unable to completely distinguish it from other templates.

The following extended generous replicator model includes this effect of partial sequence recognition, allowing for an analysis in both cases when sequences are and are not folded before they interact. In addition to the reactions given in Equation 1, in which all molecules are recognized by the catalyst X, the pool of exploiting Y molecules has to be split into two parts, namely, those (Y_0) that still have the right recognition sequence and will therefore be copied by X, and those (Y_1) that are no longer recognized by X and will not be copied. Denoting the sequence length by ν and the length of the recognition region by k, the probability that a mutation of X will produce a replicable species Y_0 depends on a function $S = S(\nu, k)$ and the model of recognition

(Sections 2.2.1, 2.2.2, 2.2.3 below). Taking the effects of replication, recognition, and decay into account, the full reaction scheme of the extended generous replicator model is

$$\begin{aligned}
\mathrm{X}+\mathrm{X}+\emptyset &\xrightarrow{Q} \mathrm{X}+\mathrm{X}+\mathrm{X},\\
\mathrm{X}+\mathrm{X}+\emptyset &\xrightarrow{(1-Q)S} \mathrm{X}+\mathrm{X}+\mathrm{Y}_0,\\
\mathrm{X}+\mathrm{X}+\emptyset &\xrightarrow{(1-Q)(1-S)} \mathrm{X}+\mathrm{X}+\mathrm{Y}_1,\\
\mathrm{X}+\mathrm{Y}_0+\emptyset &\xrightarrow{Q+(1-Q)S} \mathrm{X}+\mathrm{Y}_0+\mathrm{Y}_0,\\
\mathrm{X}+\mathrm{Y}_0+\emptyset &\xrightarrow{(1-Q)(1-S)} \mathrm{X}+\mathrm{Y}_0+\mathrm{Y}_1,\\
\mathrm{X},\mathrm{Y}_0,\mathrm{Y}_1 &\xrightarrow{d} \emptyset.
\end{aligned} \tag{2}$$

The copying quality factor Q is assumed independent of sequence type, and errors in copying recognized sequences also have a probability S of resulting in replicable Y_0. In the case of no discrimination ($S = 1$), the reactions will reduce to the ones in Equation 1 on substituting $\mathrm{Y}_0 \to \mathrm{Y}$. The stochastic transition rates for the PRESS method, deducible from Equation 2, are given in Appendix 3.

2.2.1 Linear Sequence Recognition

In the case of non-folded recognition, linear subsequences of monomers determine the recognition properties of a biopolymer. In the model, k fixed but not necessarily consecutive monomer positions are used to determine the binding region: if the recognition regions of the catalyst and target sequences match, the molecules are able to bind, so that replication can take place. In general, for sequences with a total length of ν monomers, a recognition region of k, and $m \geq 1$ point mutations, the probability for non-discrimination given mutation is

$$S_l(k) = \binom{\nu-k}{m} \Big/ \binom{\nu}{m} = \frac{(\nu-k)!(\nu-m)!}{\nu!(\nu-k-m)!} \xrightarrow[k \ll \nu]{} \left(\frac{\nu-k}{\nu}\right)^m \xrightarrow[m \approx 1]{} \frac{\nu-k}{\nu}, \tag{3}$$

that is, all mutations are located outside the recognition region. The given approximations are valid for recognition regions that are short compared with sequence length and for fidelities Q near the error threshold for which the mean number of mutations (given that mutation takes place) $m \approx 1$; refer to Appendix 1. Note that the linear approximation gives the right result, 0, for $k = \nu$. According to Equation 3, S_l will tend to 1 for $\nu \to \infty$ and constant k (that is, no further discrimination and therefore no advantage for the catalytic species).

2.2.2 Recognition with Complete Sequence Folding

Specific molecular recognition depends on a match of structural features (lock and key docking), so that major features of the sequence dependent recognition can be captured using an appropriate model of structure. This will be modified of course by induced fit. For uniquely folding biopolymers, the primary sequence of molecules determines their structure: in equilibrium, folding the final structure is the result of minimizing the free energy of alternative structures. As a result of the enormous number of combinatorial sub-optimal solutions, the calculation of the mapping from primary sequences to tertiary folded structures is extremely complex. Recent studies revealing the correlation statistics of folding, at the level of secondary structure, both for RNA [13] and for proteins [3], have shown that even a single monomer substitution can

change the resulting structure considerably. This feature is expected to be preserved in non-deterministic, non-equilibrium, and tertiary models of folding.

Folding is first treated in an extreme case of the generous replicator model (complete folding) in which each mutation results in a new random fold of the structure. Recognition still depends on a short fixed subset of positions, but now of the folded structure. The consequence of such folding is that matching sequences are now distributed over the whole structure space. For simplicity, a deterministic folding is assumed and the structure space is assumed large compared with the number of discriminated structures. A direct consequence of this scrambling between sequence and structure is that two sequences differing in only one position have major structural differences, and there is a significant probability that the recognition region will also be affected, resulting in a non-replicable sequence.

Assuming that the space of structures can be represented by sequences of length ν' over an alphabet of size κ, the probability for the folded configuration of Y having all k recognition site elements matching those in the folded structure of X is thus

$$S_f(k) = \frac{\kappa^{\nu'-k} - 1}{\kappa^{\nu'} - 1} \xrightarrow[\nu' \to \infty]{} \kappa^{-k} < 1, \tag{4}$$

yielding a finite discrimination probability even for relatively small recognition lengths k, in contrast to the linear model. In general, the folded structural motifs for recognition will be composed of a different alphabet size of structural primitives, but in order to evaluate the influence of folding under constant recognition, the alphabet size is initially taken as unaltered (cf. below).

2.2.3 Recognition Based on RNA Folding

In order to make contact with more detailed folding models of biopolymers, the data on RNA secondary structure as studied statistically [13] is utilized. The correlation between sequences and structures was studied in that work by comparing the Hamming distance between pairs of sequences with the distance between their minimal-free-energy structures using an appropriate metric. Non-knotted RNA secondary structures may be represented as trees and can be transformed into one another by a series of (tree) editing operations with predefined costs. The minimal edit distance between two structures is defined as the smallest sum of these costs along an editing path. This editing can also be regarded as a generalization of sequence alignment. For sequences with length $\nu = 100$, the secondary structure trees can be represented as strings of length 60, that is, with structure length $\nu' = 60$. Correlations in structure were found to decay rapidly and exponentially with increasing Hamming distance of the primary sequences.

Proceeding as above, a catalytic recognition domain is specified now in terms of a target substructure tree, encoded by a string of a certain length. The structure distance tends to $\nu'/2$ for the case of extremely high mutation rates, just as in the linear and complete folding modes, enabling a clear comparison of the various recognition schemes. In the following, the correlation data taken from [13] is used as summarized in Figure 1. It suffices to consider the case of only single point mutations per replication (see again the calculation given in Appendix 1).

Let a_i be the empirical probability that a single point mutation in the primary sequence leads to a structure distance i. As in Equation 3, the probability of non-

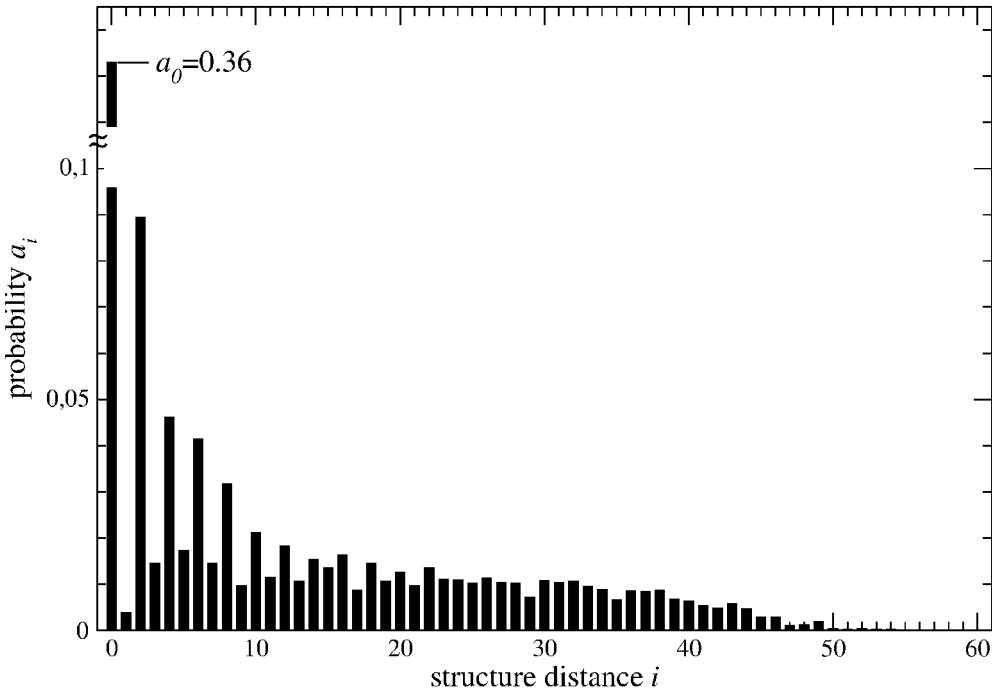

Figure 1. Single-mutation-induced structural changes in RNA. The coefficients a_i shown define the probability for a given structure distance i to be induced by a single point mutation in the primary sequence. They were extracted from the sequence-structure correlation plot calculated in [13] and are required in order to derive the recognition probability $1 - S_r$.

discrimination, in this case S_r, then becomes

$$S_r(k) = \sum_{i=0}^{\nu'} a_i \frac{\binom{\nu'-i}{k}}{\binom{\nu'}{k}} = \sum_{i=0}^{\nu'} \frac{(\nu'-k)!(\nu'-i)!}{\nu'!(\nu'-k-i)!} a_i. \tag{5}$$

Since $a_0 \neq 0$, we have $S_r = a_0 \neq 0$ even for $k = \nu'$ (in contrast to the linear and complete folding modes, where $S_{l/f} = 0$ for $k = \nu$), so that some sequence changes go unrecognized even if all structural motifs are distinguished. For RNA in solution, recognition will involve primary, secondary, and tertiary structure, allowing stronger discrimination than in the case of purely secondary structure (i.e., closer to the complete folding case). A calculation of the combined effect would involve a detailed reevaluation of the sequence-structure correlation statistics beyond the results of [13].

In summary, for each of the three recognition schemes, a probability S has been calculated that is the fraction of the Y sequence space that belongs to recognized sequences Y_0. For fixed sequence (structure) length $\nu = \nu'$, the effects of recognition under these three models of folding are compared in Figure 2, where we plot the probability $1 - S$ of a mutation event resulting in a non-replicable (i.e., "harmless") sequence, as a function of the extent of recognition k.

3 Results

To evaluate the impact of structural folding on the evolutionary stability of trans-acting catalysts, a quantitative measure of evolutionary viability is required. The maximal

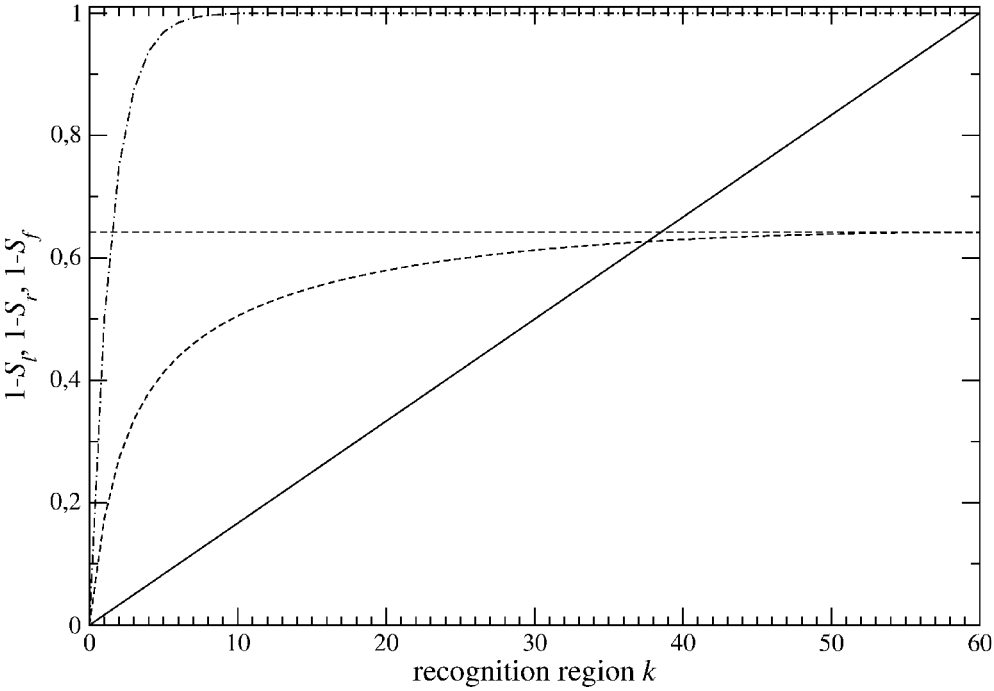

Figure 2. Discrimination probability $1 - S$ as a function of recognition length k. The graph compares the three different recognition modes: solid line for linear ($1 - S_l$), dashed line for secondary-structure-based ($1 - S_r$), and dash-dot line for the complete folding mode ($1 - S_f$). For small k, the increase of folded recognition is much higher than for unfolded, giving a significant advantage for the catalysts even for short recognition regions. Since $a_0 \neq 0$, the recognition probability $1 - S_r$ tends to $1 - a_0 \neq 1$ (horizontal dashed line) for $k \to 60$, in contrast to both the other schemes.

sustainable mutation rate for population viability provides such a measure. A critical mutation rate is well known from the quasi-species theory [8] to characterize the limits of evolutionarily stable retention of sequence information. For example, in the illustrative case of a single fastest-replicating (wild-type) sequence, its population drops to a negligible level when the overall error rate, $R = 1 - Q$, per replication exceeds a critical value R_c. A sharp error catastrophe can also be seen in the interacting systems of generous replicators studied here, but depends on the spatial diffusion coefficients and the dimensionality of the space in which the population is structured. More precisely, the limits of viability are calculated in the reduced system (see Equation 2) for various levels of sequence recognition in each of the three models introduced in the previous section. In the absence of template recognition, the maximally generous replicator model $2X \to 3X$, in which X copies any template sequence, is stable within a certain range of diffusion coefficients, as shown in [21]. With template recognition, Figure 3 shows R_c for the three different recognition modes within the stabilizing range of diffusion coefficients for $k \in \{0, \ldots, 10\}$. Enlarging the recognition region length k enhances the system's stability, and therefore the value of the critical error rate, but does not change the diffusion interval where general stability occurs. The mode employing the RNA structure correlation data lies between the two extremes of recognition, namely linear and complete folding.

Complete folding yields nearly perfect protection against replicable non-catalytic sequences (exploiters) even for short recognition lengths k. Unfolded recognition is basically worthless, because it protects only the recognition sequence itself. RNA-based

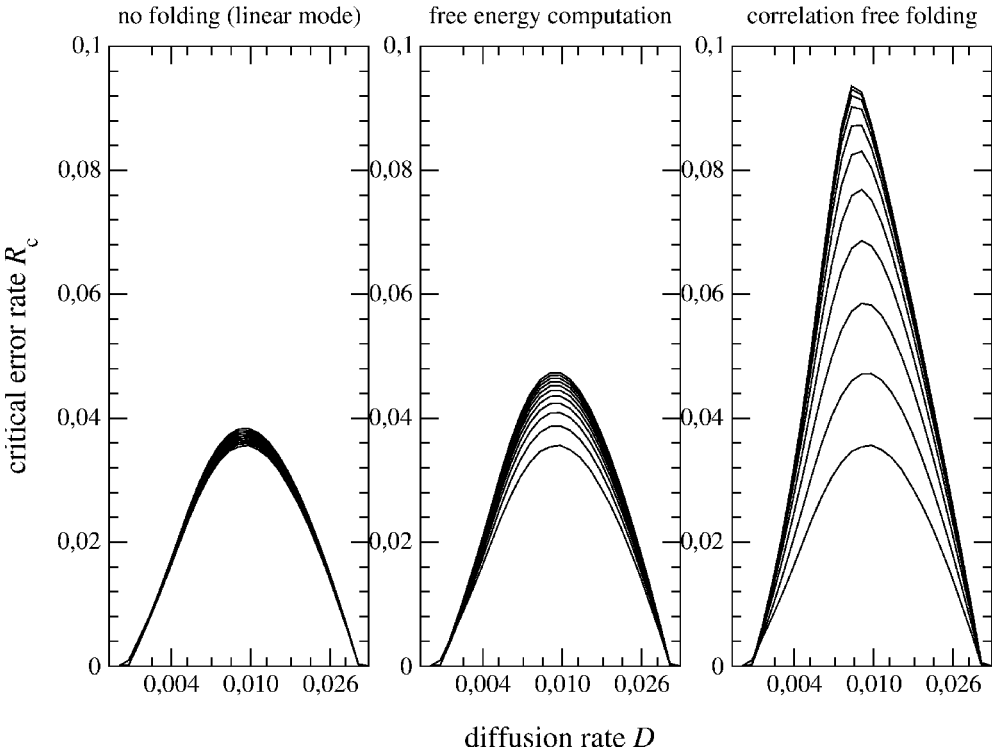

Figure 3. Critical error rate R_c as a function of the diffusion rate D. The results were calculated analytically with the PRESS platform using the three different recognition models as in Figure 2. In all cases, there is an enhancement of the system's stability for increasing recognition lengths, $k \in \{0, \ldots, 10\}$ (for each graph from bottom to top). However, a further enhancement is visible for the cases that sequence folding is taken into account, that is, for the secondary structure case (middle) as well as for the complete folding scheme (right). The $k = 0$ case, without recognition, is that studied in [21].

recognition generally does not yield complete folding or perfect protection, but even short recognition sequences do have a strong protective effect.

The PRESS framework permits an extrapolation to the effective behavior in three and lower dimensional space, as shown in [21]. Because the accuracy of this reduction may be model dependent, a similar comparison of the predictions of the dimensionality-extrapolated PRESS model with the results of Monte Carlo simulations for the kinetic scheme of Equation 2 is presented in Figure 4 for linear and complete sequence recognition (both evaluated for $k/\nu = 0.1$, i.e., a 10% recognition length). These stochastic simulations were calculated for the full sequence space resolution (beyond the reduced three species model) with an overall population of 120,000 binary strings, each having a length $\nu = 20$. For further information on the Monte Carlo simulations in connection with the PRESS model see also [2]. The simulation results show, for the infinite dimensional case, good agreement with the PRESS results (see Figure 4), which not only confirms the viability of the PRESS approach but also verifies the three species model employing the parameter S. For the simulations in lower dimensional Euclidean spaces (retaining full sequence resolution), the tolerable error rates, and so also the lengths of sequences which can be functionally stabilized by limited recognition, are significantly larger in absolute terms.

Since the length of encoded information is limited by catalytic precision, one is curious to see if the length of the coding sequences can be enlarged indefinitely by the

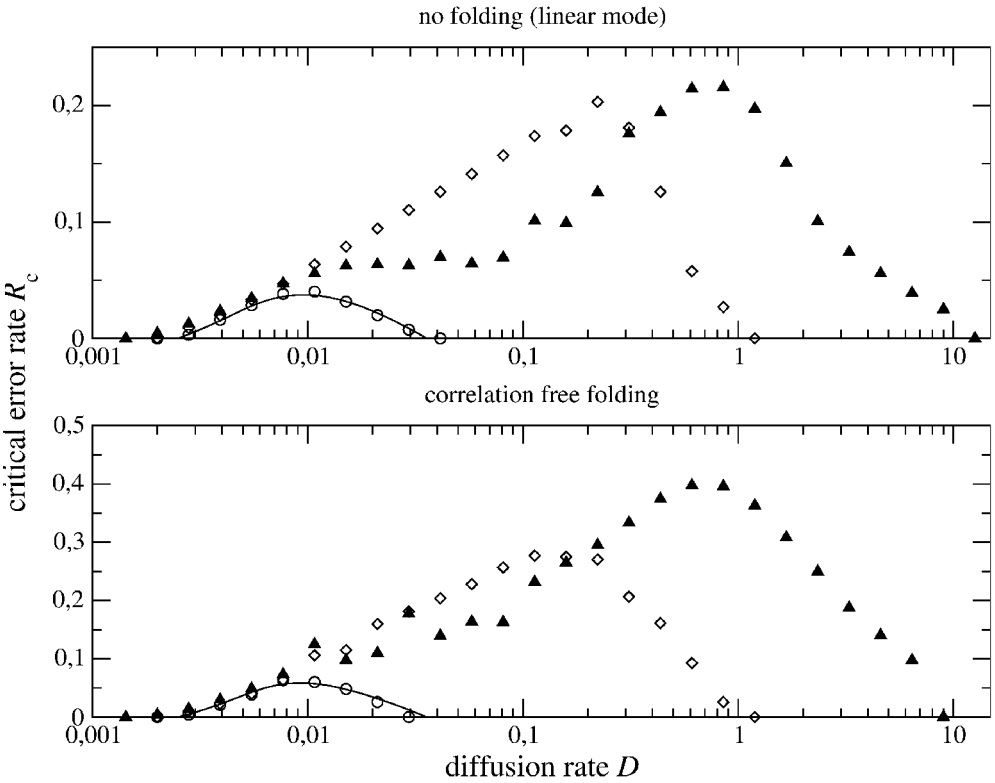

Figure 4. Critical error rate R_c: results of the stochastic simulations with full sequence resolution for non-folded (upper graph) and completely folded (zero correlation, lower graph) recognition. The different symbols correspond to the different physical dimensions: ○ for the infinite, ◇ for the three, and ▲ for the two dimensional case. The solid line corresponds to the PRESS results for the three-species model and shows good agreement with the stochastic simulations. For every simulation, a constant recognition region length of 2 (10% of the whole sequence length) was used. The results show significantly enhanced stabilization over those without recognition, reported in [21].

effects of recognition. To test this, functionally redundant recognition elements were added consecutively to the end of the coding region. The question is then whether the stabilizing effects of recognition allow room for a further extension of the coding region under the condition that the fidelity of monomer replication does not change.

The PRESS model allows one to calculate R_c, which depends on the recognition model and the length k of the recognition region. Assuming constant chemical conditions, implying a constant monomer replication fidelity q, there is a simple relationship between the maximal sustainable sequence length ν and the critical error rate R_c, namely, $R_c = 1 - q^{\nu}$. To calculate the appropriate error rates within the PRESS model, the recognition parameters S in the three cases above have to be rearranged for $\nu \to \nu + k$. As one increases the length of the recognition region from zero, R_c increases, reflecting enhanced system stability. According to the relation above, this additional stability can be used to increase the sequence length ν. However, the addition of recognition monomers has already increased the sequence length somewhat, and so, can additional monomers beyond the recognition monomers be gained? It turns out that the answer to this question depends on the recognition model as shown in Figure 5, where the sequence length extension ratio ε is shown. Only for the folded models of recognition are additional monomers that can be used to encode improved catalysts gained ($\varepsilon > 1$). This stabilization is seen to be large even for very few recog-

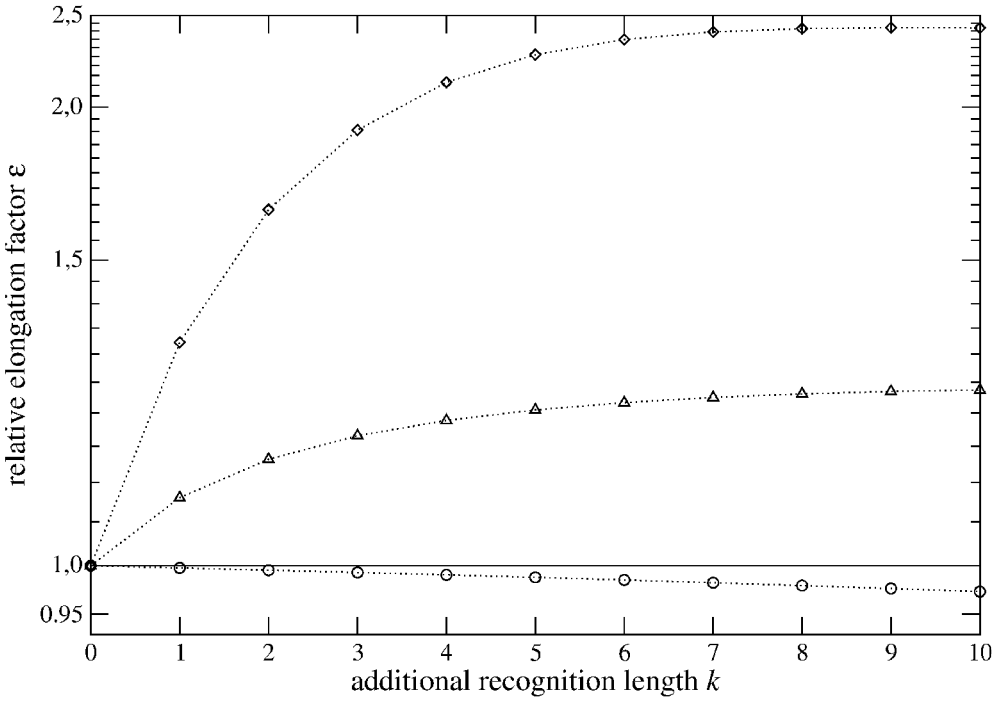

Figure 5. Useful sequence enlargement ratio for the three different recognition modes. For an error rate that limits sequence length to $\nu = 100$ in the absence of recognition, the possible enlargement ratio $\varepsilon(k)$ for sequence elements not involved in recognition coding is plotted as a function of the additional recognition length k. Only when sequence folding modulates recognition ($\diamond$ for correlation free folding and $\triangle$ for the free energy computation) can the additional stability be used to enlarge the sequences up to the plotted values of ε. For the linear case ($\circ$), no enlargement is possible, that is, $\varepsilon < 1$ (see text).

nition monomers. The additional stability could of course also be used to allow higher error rates.

Although the results shown in Figure 5 have been calculated for a fixed monomer fidelity q, the value at the threshold of stability for $\nu = 100$ without recognition, one can scale ε for various different sequence lengths ν and appropriate k, making use of the fact that the critical error rate is simply a function of the non-discrimination probability S, $R_c = R_c(S)$, which can be determined by numerical solution of the closed nonlinear master equation in the PRESS formulation or determined directly from Monte Carlo simulations. In particular, in the case of the complete folding mode, ε will not change if k is fixed, provided ν is large enough so that S_f becomes independent of ν, as can be seen from Equation 4. This constant relative enhancement of useful sequence length, independent of ν, demonstrates that the complete folding scheme will always lead to a clear overcompensation. Of course, this is just the effect of recognition alone and does not imply that other selective properties are constant as sequence length increases. In the unfolded recognition scheme, on the other hand, the additional recognition monomers can never overcompensate their own induced sequence enlargement.

4 Discussion

The above results demonstrate the impact of recognition, (i.e., molecular discrimination) on the evolutionary stability of trans catalysis. Recognition was seen to allow an increase of stability for the catalytic system of the generous replicator, measured quantitatively

in terms of the maximal sustainable error rate. Folding based models of molecular recognition show that even quite limited local recognition can be effective in evolving stable catalysts in spatially resolved systems. Using the PRESS model, this result was analytically proven for infinite dimensional systems for a reduced three species system. Monte Carlo simulation, verifying these results also for the case of a full sequence-space resolution, showing additionally that for two and three dimensions this stability increment is even stronger in absolute terms.

Subsequent calculations allowed an estimation of whether additional monomers enhancing recognition but increasing the total sequence length provide a net advantage for trans-catalytic activity. This calculation showed the fundamental difference between linear and folded recognition. For the linear case, the additional monomers were at best just able to compensate their own cost in contrast to the complete folding model, where a clear overcompensation was detected. This overcompensation is the result of the scrambling effects of folding that spread the recognized sequences over sequence space, achieving a great enhancement of stability without enlarging the recognition region length.

Note that we have adopted a rather arbitrary unit for quantifying the complex phenomenon of molecular recognition in this work, assuming that it can be reduced to linear pattern recognition between words over a finite alphabet. We assume here that recognition is local and make the conservative assumption that the number of structural patterns that can be distinguished locally is not increased significantly by folding. In the case of RNA secondary structure studied in [13], it is in fact reduced from 100 to 60 if we ignore individual sequence recognition, but increased from 100 to 160 if we include it. This of course would further enhance the ability of folding to stabilize trans-catalytic cooperation. This work can be understood as investigating the pure influence of information folding rather than including additional structural differentiation, which certainly further enhances the ability to generate sequence information.

These results outline analytically the fundamental role of folding, beyond straightforward functional determination, in the context of molecular evolution. Moreover, they demonstrate that full and therefore unphysical sequence recognition, as implicit in the ideal hypercycle theory [9], is not necessary in order to stabilize trans-catalytic reaction systems. The small advantage of catalytic species resulting from spatial resolution provides a sufficient platform on which partial sequence recognition can build further stability. Recognition by folding overcompensates the costs of its implementation. This opens the possibility of neutral evolution, with respect to overall system stability, wherein the destabilizing effects of catalyst specification by sequence enlargement are compensated by the simultaneous addition of partial recognition machinery.

Using the PRESS method, the efficacy of correlation based maintenance of trans-catalytic activity in the generous replicator model was demonstrated [21], without the need for precise recognition [9], self-replicating compartments [23], or specific spatial structures such as spirals [4]. Appropriate limited recognition based on folding greatly enhances this capability, as described above. Among other different models that allow the study of spatially resolved systems it is worthwhile to compare it with the well-known metapopulation methods of population biology [18]; for a recent review see [12]. The PRESS method differs from the metapopulation approach in several aspects; one of the major ones is that the PRESS method takes the internal dynamics of a site explicitly into account. Apart from the stated advantages of a detailed description of discrete small populations, this leads to an extinction-growth behavior that does not have to be postulated by phenomenological reasoning, but is the generic result of reaction equations and self-organization, which in turn means that the time evolution of the system can be calculated ab initio from experimentally accessible reaction rates.

Molecular recognition and discrimination as a method to stabilize "altruistic" trans catalysts finds a parallel in kin selection theory [11], as mentioned previously. The notion of inclusive fitness [1], one of the basic ideas in kin selection theory, also can be applied to trans catalysts as "altruists" that can only survive within a group of relatives. The familiar problem in kin selection theory, namely the recognition of genetically related individuals beyond simple brood colocalization, must also be solved at the molecular level. Since it is impossible to discriminate between all changes in the genome, purely molecular kin recognition can never be complete and must be complemented by spatial correlations. In higher organisms, a wealth of basic genetic functionalities, such as the calls or visual displays of birds or the scents of animals, are often recognized in "folded" form by means of a few complex tests [14, 22]. Of course, we do not want to overstate the relevance of our results to higher organisms, merely pointing out that the degree of convolution (folding) of the dependence of phenotype on genotype has a significant impact on recognition based selection.

As indicated in previous work [2, 21], the range of diffusion constants relevant for trans-catalytic stabilization is both reasonable in thin films at the origin of life and accessible to molecular biotechnology in vitro [10]. In further studies one has to show that these results can be transferred to experimental evolution in vitro, for example, to RNA-based replication systems. Finally, further applications may be found in evolutionary molecular engineering and self-organized supramolecular chemistry [17].

Acknowledgments

This work was partially supported by the Deutsche Forschungsgemeinschaft (DFG) and the Swiss National Science Foundation (SNF).

References

1. Alonso, W. J. (1998). The role of kin selection theory on the explanation of biological altruism: A critical review. *Journal of Comparative Biology, 3*, 1–14.
2. Altmeyer, S., & McCaskill, J. S. (2001). Error threshold for spatially resolved evolution in the quasispecies model. *Physical Review Letters, 86*, 5819–5822.
3. Babajide, A., Farber, R., Hofacker, I. L., Inman, J., Lapedes, A.S., & Stadler, P.F. (2001). Exploring protein sequence space using knowledge based potentials. *Journal of Theoretical Biology, 212*, 35–46.
4. Boerlijst, M. C., & Hogeweg, P. (1991). Spiral wave structure in pre-biotic evolution-hypercycle stable against parasites. *Physica D, 48*, 17–28.
5. Bresch, C., Niesert, U., & Harnasch, D. (1980). Hypercycles, parasites, and packages. *Journal of Theoretical Biology, 85*, 399–405.
6. Cronhjort, M. B., & Blomberg, C. (1997). Cluster compartmentalization may provide resistance to parasites for catalytic networks. *Physica D, 101*, 289–298.
7. Eigen, M. (1971). Self-organization of matter and evolution of biological macromolecules. *Naturwissenschaften, 58*, 465–523.
8. Eigen, M., McCaskill, J. S., & Schuster, P. (1989). The molecular quasi-species. *Advances in Chemical Physics, 75*, 149–263.
9. Eigen, M., & Schuster, P. (1977). Hypercycle—Principle of natural self organization. *Naturwissenschaften, 64*, 541–565.
10. Fritz, J., Baller, M. K., Lang, H. P., Rothuizen, H., Vettiger, P., Meyer, E., Güntherodt, H.-J., Gerber, C., & Gimzewski, J.K. (2000). Translating biomolecular recognition into nanomechanics. *Science, 288*, 316–318.
11. Hamilton, W. D. (1964). The genetical evolution of social behavior. *Journal of Theoretical Biology, 7*, 1–52.

12. Harding, K. C., & McNamara, J. M. (2002). A unifying framework for metapopulation dynamics. *American Naturalist, 160*, 173–185.

13. Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, S., Tacker, M., & Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatsheft für Chemie, 125*, 167–188.

14. Isles, A. R., Baum, M. J., Ma, D., Keverne, E. B., & Allen, N. D. (2001). Gene imprinting: Urinary odor preference in mice. *Nature, 409*, 783–784.

15. Kauffman, S. A. (1971). Gene regulation networks: A theory of their global structure and behavior. *Current Topics in Developmental Biology, 6*, 145–182.

16. Kimura, M. (1983). Diffusion model of intergroup selection with special reference to evolution of an altruistic character. *Proceedings of the National Academy of Sciences of the U.S.A., 80*, 6317–6321.

17. Lehn, J. M. (2002). Toward complex matter: Supramolecular chemistry and self-organization. *Proceedings of the National Academy of Sciences of the U.S.A., 99*, 4763–4768.

18. Levin, S. A., & Paine, R. T. (1974). Disturbance, patch formation and community structure. *Proceedings of the National Academy of Sciences of the U.S.A., 71*, 2744–2747.

19. McCaskill, J. S. (1994). "Inaugural lecture at the Friedrich Schiller University, Jena, Germany." In *Urspruenge der molekularen Kooperation: Theorie und Experiment.* Jena, Germany: IMB Press.

20. McCaskill, J. S. (1997). Spatially resolved molecular in-vitro ecology. *Biophysical Chemistry, 66*, 145–158.

21. McCaskill, J. S., Füchslin, R. M., & Altmeyer, S. (2001). The stochastic evolution of catalysts in spatially resolved molecular systems. *Biophysical Chemistry, 382*, 1343–1363.

22. Parr, L. A., & de Waal, F. B. (1999). Visual kin recognition in chimpanzees. *Nature, 399*, 647–648.

23. Szathmary, E., & Demeter, L. (1987). Group selection of early replicators and the origin of life. *Journal of Theoretical Biology, 128*, 463–486.

24. Wright, S. (1977). *Evolution and the genetics of population, Vol. 3*, Chicago: University of Chicago Press.

Appendix 1: Relative Frequency of Multiple Mutations Near the Error Threshold

A straightforward specific result needed on mutation probabilities for single monomer substitutions is derived in this appendix. For a single replication event, with q being the probability that a single monomer will be copied without error, the probability of having exactly m point mutations is given by the binomial distribution

$$P(m) = \binom{\nu}{m} q^{\nu-m}(1-q)^m, \tag{6}$$

with $P(m=0) = q^\nu = Q$. Equation 6 allows one to calculate the mean number of point mutations under the condition that the replication was not error free:

$$\bar{m} = \frac{\sum_{m=1}^{\nu} mP(m)}{\sum_{m=1}^{\nu} P(m)} = \nu\frac{1-q}{1-Q} \approx 1 + (1-q)\frac{\nu-1}{2}. \tag{7}$$

For realistic $q \approx 1$ above the error threshold, where the simulations show that typically $Q > 1/2$, the value of $\bar{m}$ is only slightly different from 1. That means that if a replication had an error, then in most cases there has been just one point mutation.

Multiple mutations are significant for evolutionary innovation, but not so important for characterizing the general problem of exploitation.

Appendix 2: Outline of the PRESS Modeling Framework

The PRESS model allows a reduced treatment of discrete particle reaction-diffusion systems. It is set up in a high dimensional space with an infinite number of local sites, each of which can contain at most a finite number of molecules, n. This model can later be projected onto Euclidean space. Chemical reactions only take place within each site. Additionally, molecular interchange is possible with a rate coefficient D_∞, where the symbol ∞ denotes the high dimensional simplex topology, entailing that each site is a neighbor to every other.

For s different chemical species, the time-dependent probabilities $P_i^\eta(t)$ of the $N_Z = (n+s)!/(n!s!)$ different composition states i at a particular site η may be collected in the probability vector $\vec{P}^\eta(t) = (P_1^\eta(t), P_2^\eta(t), \ldots, P_{N_Z}^\eta(t))^T$. Since all of the sites are statistically equivalent given symmetric initial conditions, even though individual realizations differ, we have $\vec{P}^\eta(t) = \vec{P}(t)\ \forall\eta$. In fact, this is usually also true after a limited induction period even for heterogeneous initial conditions, since the diffusive coupling dynamics involve the mean site occupation probabilities. To evaluate the time dependence of the state probability, one has to construct the transition rates from the given set of reactions. Let $s = 2$, so that only two different species X and Y are used, and let x and y be the instantaneous numbers of X and Y at a chosen site.

A typical reaction, in which X replicates Y at site η with a reaction rate coefficient α, has a discrete transition rate

$$w_{x\to x,y\to y+1} = \alpha \frac{x}{n}\frac{y}{n-1}\frac{n-x-y}{n-2},$$

where the reaction can only take place with a vacant location for the product at the site. Additionally, the site interchange processes occur with transition rates such as

$$w_{x\to x-1,y\to y+1} = D_\infty \frac{x}{n}\frac{\bar{y}}{n},$$

where for example an X is interchanged with a Y from another site. The variable $\bar{y}$ is the mean number of Y in the whole system. Due to the complete symmetry, the mean numbers of molecules of any type Z (e.g., $\bar{x}$ and $\bar{y}$) can be calculated as

$$\bar{z}(t) = \sum_{i=1}^{N_Z} P_i(t) K_i^z,$$

where K_i^z determines the number of Z in occupation state i. For every species involved, such a self-consistency equation is determined.

Taking all rates $w_{x\to x',y\to y'}$ into account in the transition rate matrix L, the time evolution of the probability vector $\vec{P}(t)$ is

$$\frac{d\vec{P}(t)}{dt} = L\vec{P}(t).$$

It is important to stress that, due to the dependence of the migration rates on the probabilities $P_i(t)$, the matrix L itself depends on $\vec{P}(t)$, that is, $L = L(\vec{P}(t))$, making the dynamics nonlinear.

In the stationary state, $\vec{P}(t) = \vec{P}$, the self-consistency equations can be used to solve for the mean population numbers and to analyze their dependence on the given parameters. For the general time dependence, the ordinary differential equations can be solved numerically.

Appendix 3: Transition Rates for the Recognition Model

The rates that are necessary to construct the matrix L are given for the three species recognition model (X, Y_0, Y_1). They can be constructed with the help of Equation 2 and take the form

$$
\begin{aligned}
w_{x\to x+1} &= Q\frac{x(x-1)z}{n(n-1)(n-2)},\\
w_{y_0\to y_0+1} &= \frac{x}{n}\left(Q\frac{y_0}{n-1} + (1-Q)S\left(\frac{x-1}{n-1}+\frac{y_0}{n-1}\right)\right)\frac{z}{n-2},\\
w_{y_1\to y_1+1} &= (1-Q)(1-S)\frac{x}{n}\left(\frac{x-1}{n-1}+\frac{y_0}{n-1}\right)\frac{z}{n-2},\\
w_{a\to a-1} &= d\frac{a}{n} + D_\infty\frac{a}{n}\frac{\bar{z}}{n}, \quad w_{a\to a+1} = D_\infty\frac{\bar{a}}{n}\frac{z}{n},\\
w_{a\to a-1,b\to b+1} &= D_\infty\frac{a}{n}\frac{\bar{b}}{n}, \quad a, b \in \{x, y_0, y_1\}, \quad a \neq b, \quad z = n-(x+y_0+y_1),
\end{aligned}
$$

where $\bar{x}$, $\bar{y}_0$, and $\bar{y}_1$ are the average values of these variables over all sites, computed self-consistently from the resulting probability distributions (cf. [21]). Depending on the choice of folding scheme, the parameter S is replaced by S_l, S_r, or S_f. We have assumed a constant diffusion rate for all species.