מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

Compositional Assemblies Behave Similarly to Quasispecies Model

Renan Gross, Omer Markovitch and Doron Lancet

Department of Molecular Genetics, Weizmann Institute of Science, Israel

Group meeting, 01/10/2013

Introduction

Quasispecies are a cloud of genotypes that appear in a population at mutation-selection balance.

(J.J Bull et al, *PLoS 2005)*

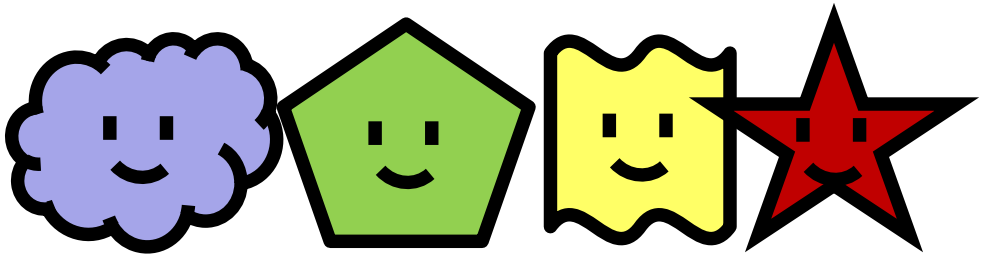
- Theoretical model (equations and assumptions), with experimental support by RNA viruses.
- Usually applied when mutation rates are high.

- GARD composomes replicate with relatively low fidelity (high mutation rate).

Could they show similar dynamic behavior?

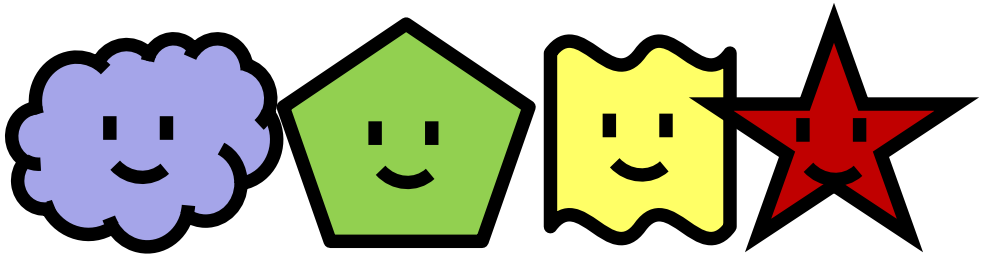
Quasispecies model

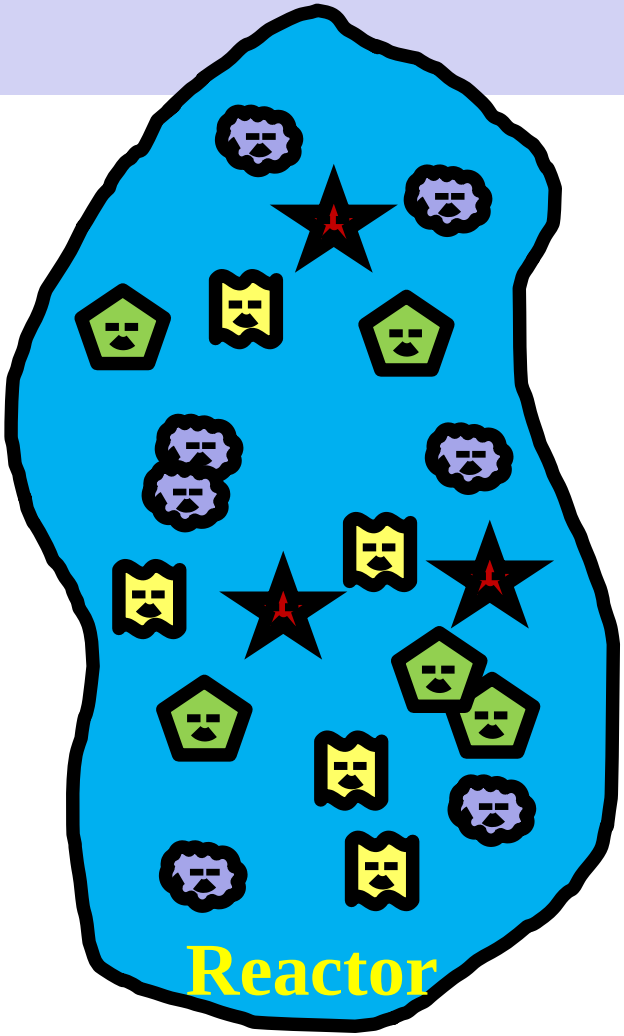
- Basically a population model
- *n* different genotypes / identities



Quasispecies model

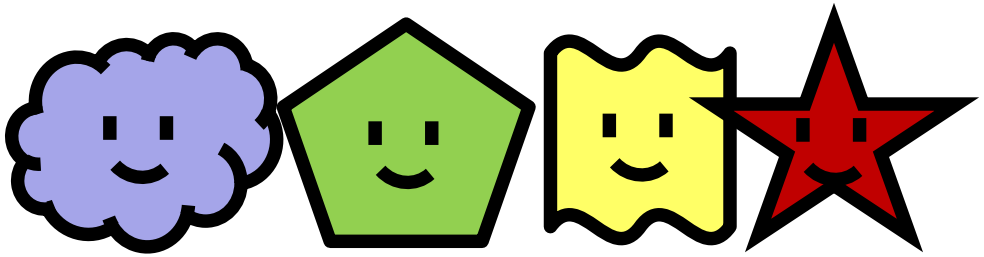
- Basically a population model
- *n* different genotypes / identities



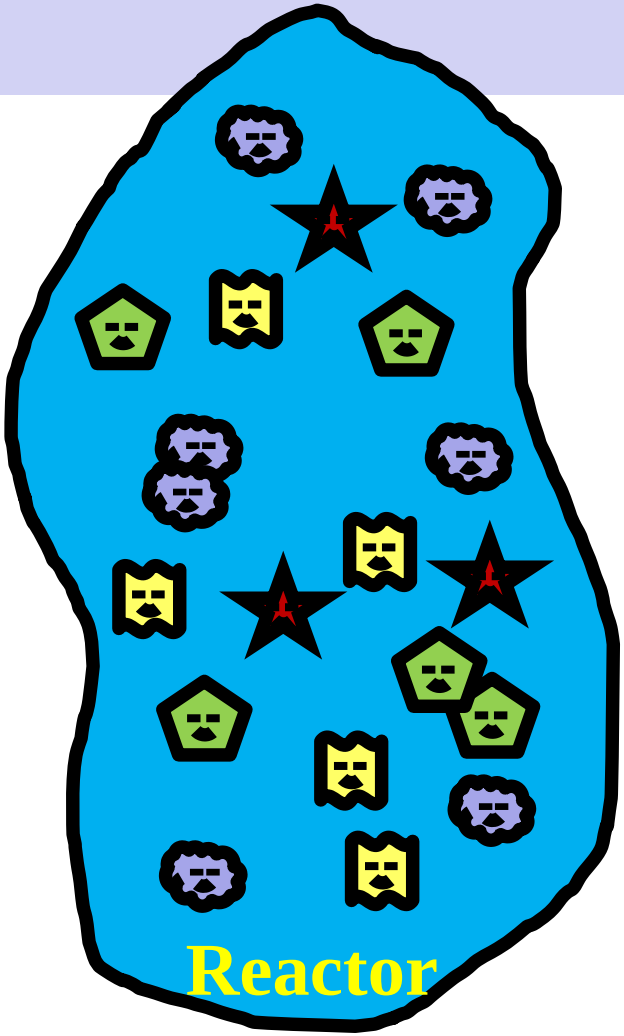


Quasispecies model

- Basically a population model
- *n* different genotypes / identities



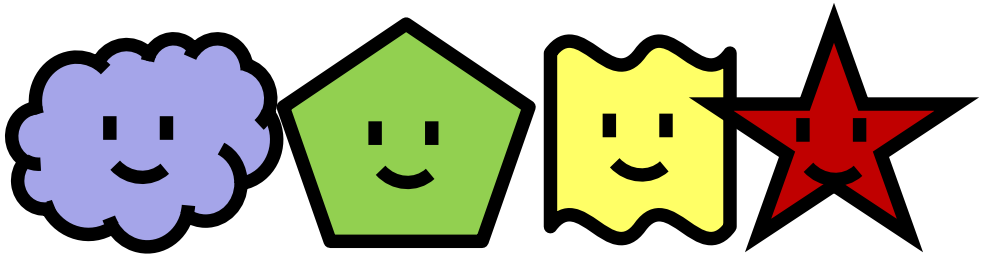
- Their relative **concentration** in the environment is denoted by the fraction x_i.

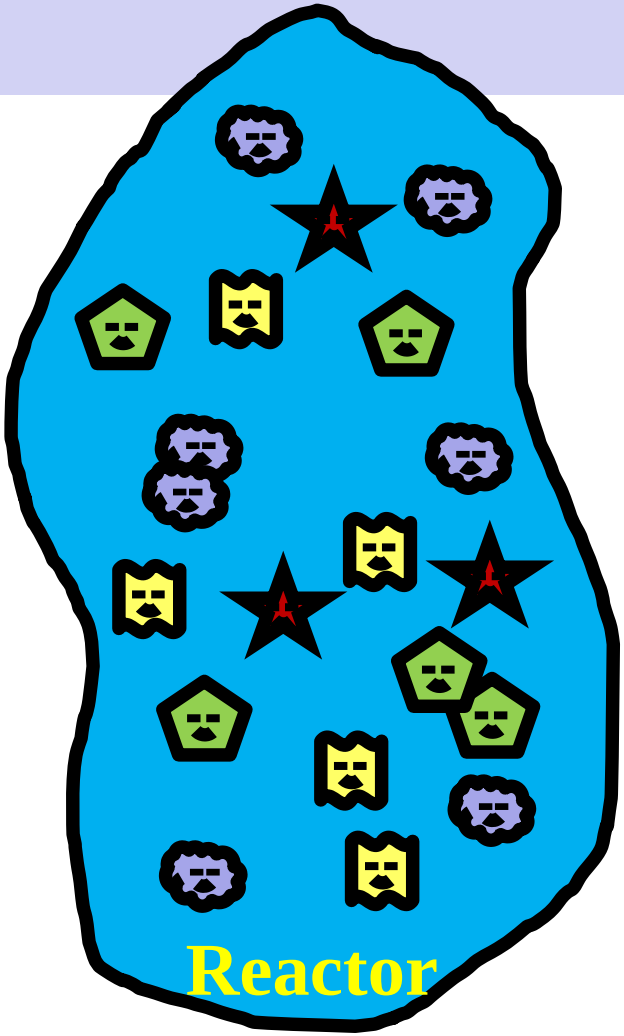


$$\sum_{i=1}^{n} x_i = 1$$

Quasispecies model

- Basically a population model
- *n* different genotypes / identities



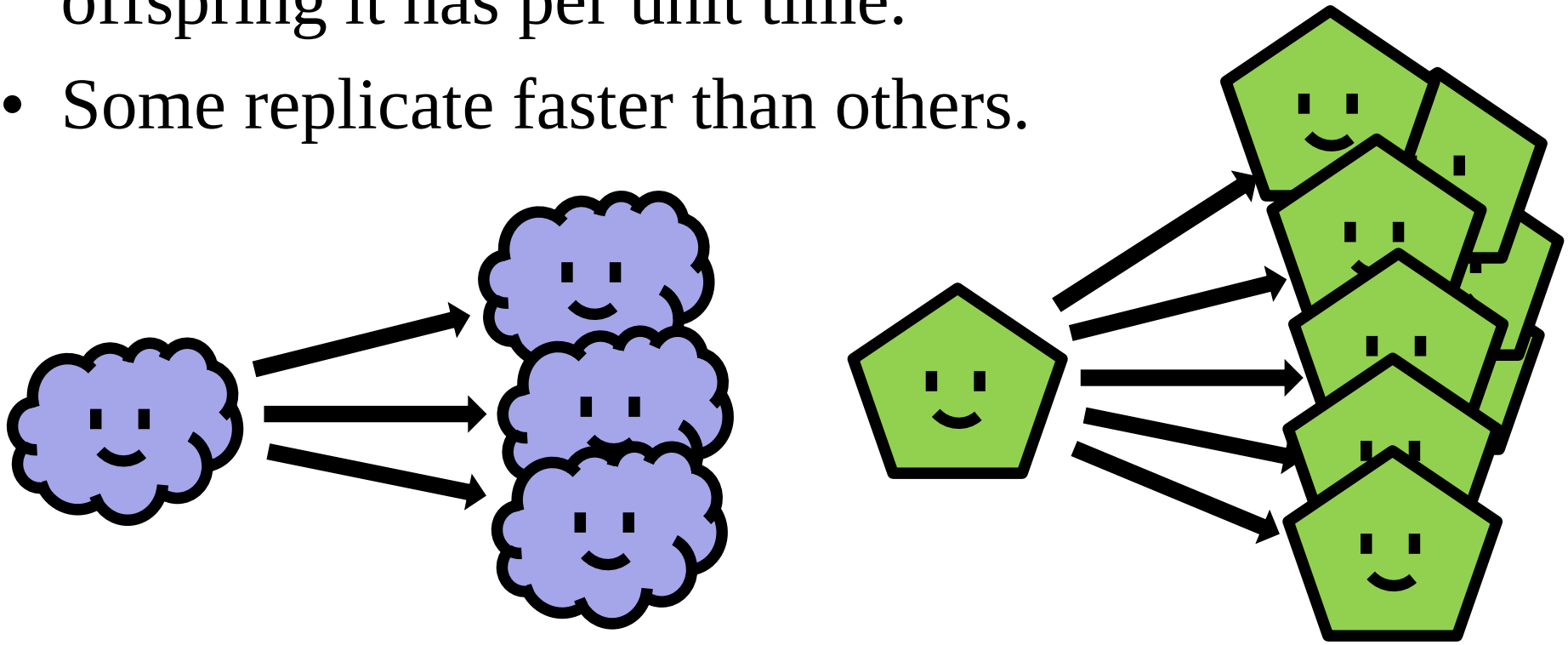


- Their relative **concentration** in the environment is denoted by the fraction x_i.
- **Goal: To find out how x_i behave as a function of time.**

$$\sum_{i=1}^{n} x_i = 1$$

Quasispecies model

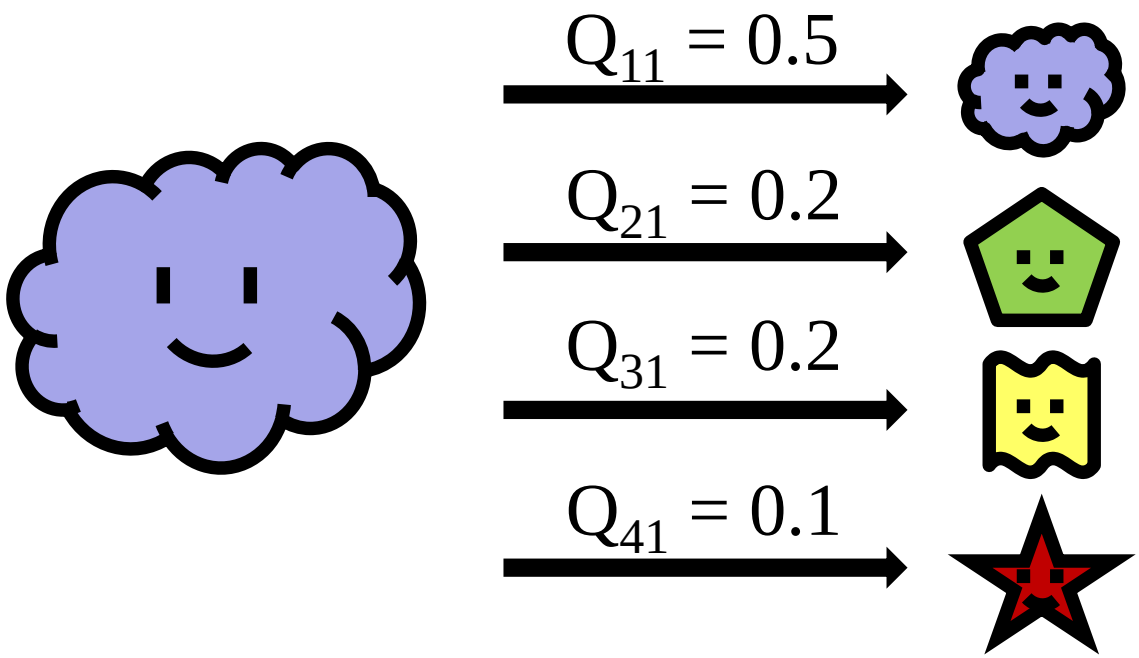
- Each one replicates at a certain rate – how many offspring it has per unit time.
- Some replicate faster than others.



- This is called the **replication rate**, denoted A_i
- $A_i = \frac{number\ of\ offspring}{time}$

Quasispecies model

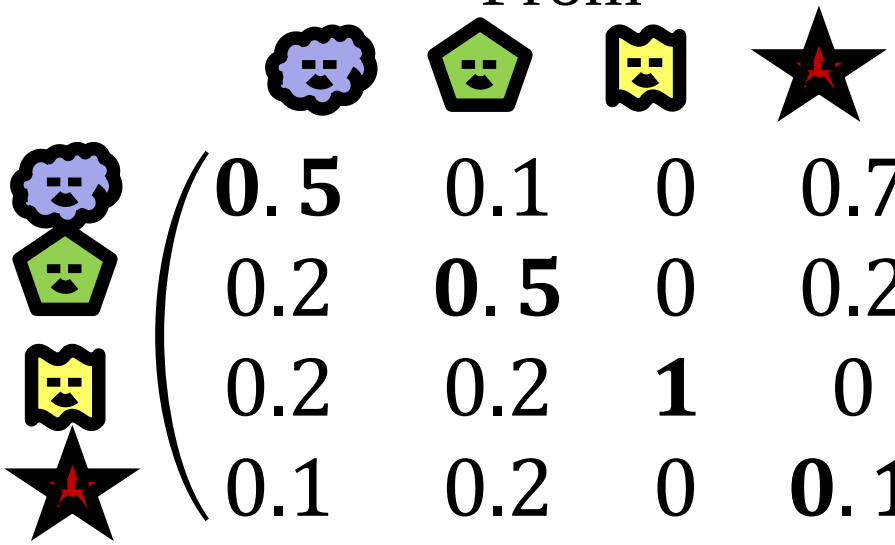
- However, replication is not exact. Sometimes, the offspring is of another genotype.
- The chance that a genotype *j* replicates into genotype *i* is denoted Q_{ij}.



Quasispecies model

- We can put everything in a matrix, called the **transition matrix, Q.**

From



To

$$\begin{pmatrix} \mathbf{0.5} & 0.1 & 0 & 0.7 \\ 0.2 & \mathbf{0.5} & 0 & 0.2 \\ 0.2 & 0.2 & \mathbf{1} & 0 \\ 0.1 & 0.2 & 0 & \mathbf{0.1} \end{pmatrix}$$

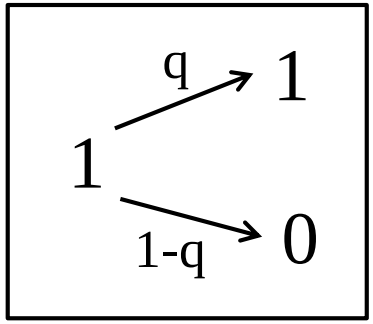
- The main diagonal is faithful self replication.

Quasispecies model

- How can we find out **Q** and **A**?
- Assume that we are dealing with RNA sequences of length *v*:
  - There is a single genotype with highest replication rate: the **master sequence**

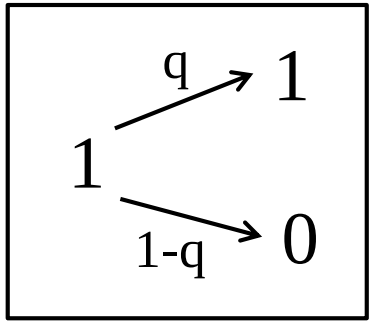
Quasispecies model

- How can we find out **Q** and **A**?
- Assume that we are dealing with RNA sequences of length *v*:
  - There is a single genotype with highest replication rate: the **master sequence**
  - Single digit replication: $0 \leq q \leq 1$
    - What is the chance for error-less replication?



Quasispecies model

- How can we find out **Q** and **A**?
- Assume that we are dealing with RNA sequences of length *v*:
  - There is a single genotype with highest replication rate: the **master sequence**
  - Single digit replication: $0 \leq q \leq 1$
    - What is the chance for error-less replication?
  - All mutations of the master sequence replicate slower according to the Hamming distance.
    - Effectively, many genotypes are grouped together.
    - **Q** and **A** are built only from *q* and *v*.



Quasispecies model

Quasispecies model

- $q = 1$ ➔

Quasispecies model

- $q = 1$ ➔ exact replication

Quasispecies model

- $q = 1$ ➔ exact replication
- $q = 0$ ➔

Quasispecies model

- $q = 1$ ➔ exact replication
- $q = 0$ ➔ exact complimentary replication

Quasispecies model

- $q = 1$ ➔ exact replication
- $q = 0$ ➔ exact complimentary replication
- $q = 0.5$ ➔

Quasispecies model

- $q = 1$ ➔ exact replication
- $q = 0$ ➔ exact complimentary replication
- $q = 0.5$ ➔ all data is lost.

Quasispecies model

- $q = 1$ ➔ exact replication
- $q = 0$ ➔ exact complimentary replication
- $q = 0.5$ ➔ all data is lost.

- Starting at $q = 1$, lowering it results in loss of the master sequence as the most frequent genotype.
  - (requires lack of back mutation)

ERROR CATASTROPHE

- RNA viruses may be fought by bringing them to error catastrophe.

Quasispecies model

- Under constant population assumptions, the transition matrix **Q** and the replication rates **A** are all that are needed in order to find out the concentration dynamics.

Quasispecies model

- Under constant population assumptions, the transition matrix **Q** and the replication rates **A** are all that are needed in order to find out the concentration dynamics.
- The *Eigen* equation (after Manfred Eigen):

$$\frac{dx_i}{dt} = \left(A_i Q_{ii} - \tilde{E}(t)\right) x_i + \sum_{j \neq i} A_i Q_{ij} x_j$$

Quasispecies model

- Under constant population assumptions, the transition matrix **Q** and the replication rates **A** are all that are needed in order to find out the concentration dynamics.
- The *Eigen* equation (after Manfred Eigen):

$$\frac{dx_i}{dt} = \left(A_i Q_{ii} - \tilde{E}(t)\right) x_i + \sum_{j \neq i} A_i Q_{ij} x_j$$

- Where E is the "**average excess rate**"

$$\tilde{E}(t) = \sum_{i=1}^{n} A_i x_i$$

Quasispecies model

- **Q** and **A** can be combined into one matrix **W**, which tells how much of each genotype is produced per unit time.

$$\boldsymbol{W} = \boldsymbol{Q} \cdot diag(\boldsymbol{A})$$

Examples:

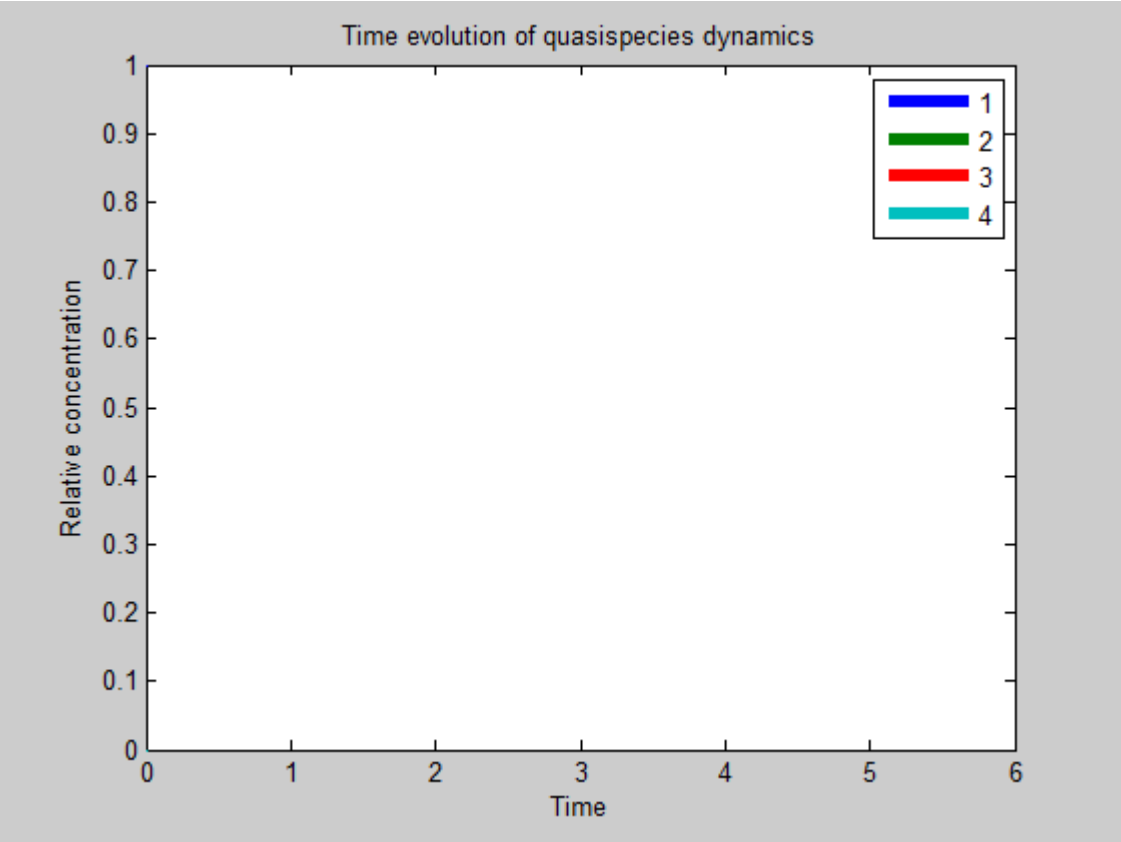
$$\boldsymbol{W} = \begin{pmatrix} 1 & 0.001 & 0.001 & 0.01 \\ 0.1 & 2 & 0.01 & 0.01 \\ 0.001 & 0.1 & 3 & 0.01 \\ 0.001 & 0.001 & 0.001 & 4 \end{pmatrix}$$

- Initial conditions: $x_1 = 1$, all others are 0.

Examples:

$$\boldsymbol{W} = \begin{pmatrix} 1 & 0.001 & 0.001 & 0.01 \\ 0.1 & 2 & 0.01 & 0.01 \\ 0.001 & 0.1 & 3 & 0.01 \\ 0.001 & 0.001 & 0.001 & 4 \end{pmatrix}$$

- Initial conditions: $x_1 = 1$, all others are 0.



Examples:

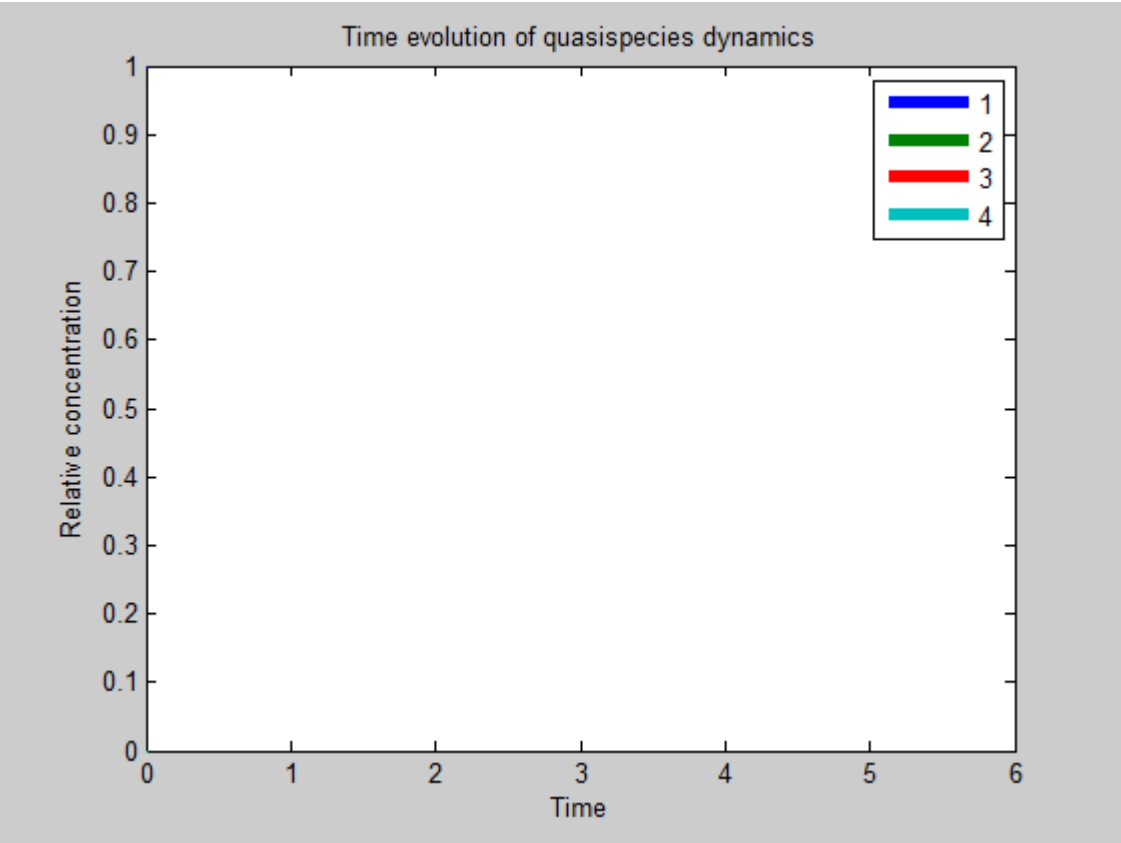
$$\boldsymbol{W} = \begin{pmatrix} 1 & 0.001 & 0.001 & 1 \\ 0.1 & 2 & 0.01 & 1 \\ 0.001 & 0.1 & 3 & 1 \\ 0.001 & 0.001 & 0.001 & 4 \end{pmatrix}$$

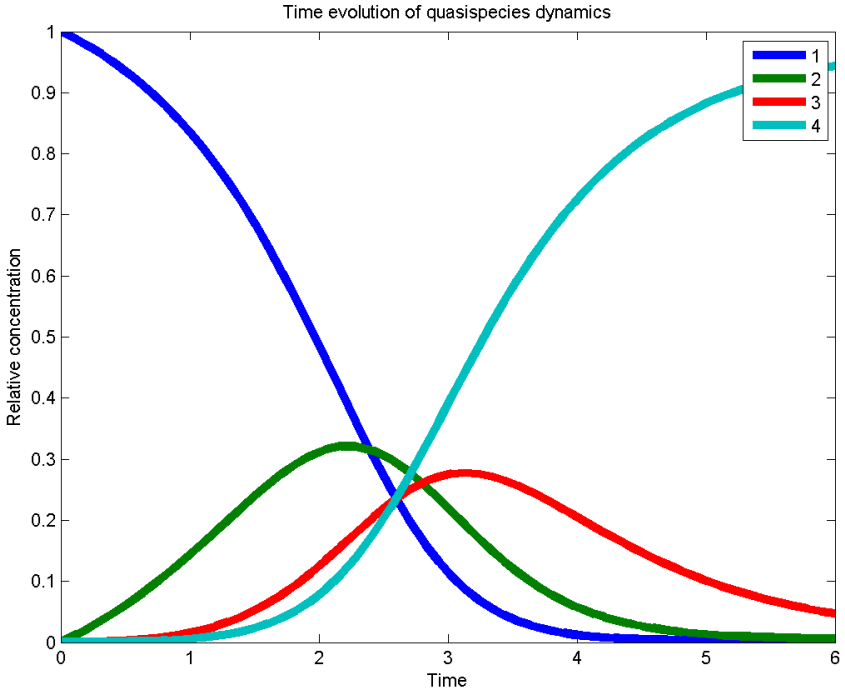
- Initial conditions: $x_1 = 1$, all others are 0.

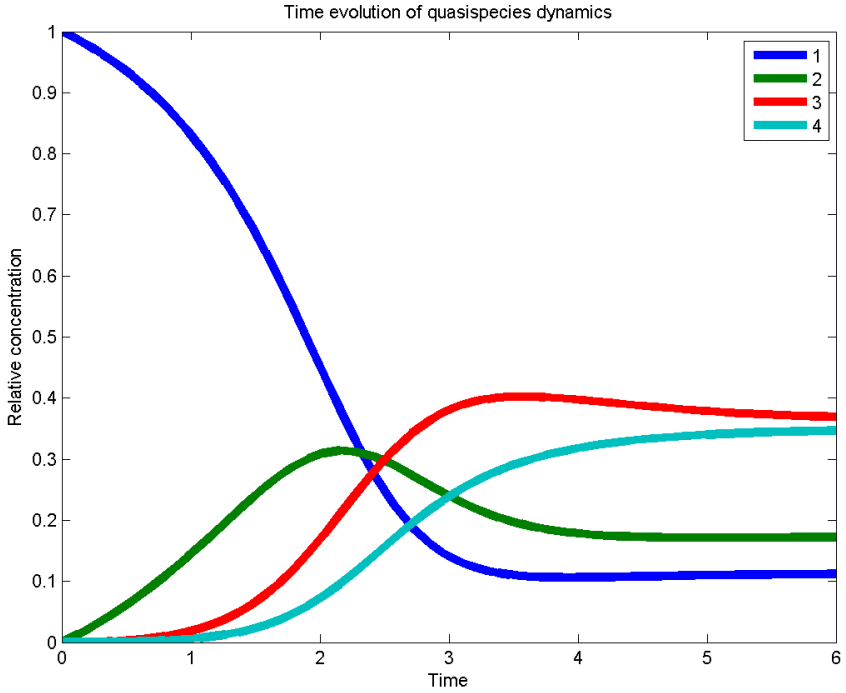
Examples:

$$\boldsymbol{W} = \begin{pmatrix} 1 & 0.001 & 0.001 & 1 \\ 0.1 & 2 & 0.01 & 1 \\ 0.001 & 0.1 & 3 & 1 \\ 0.001 & 0.001 & 0.001 & 4 \end{pmatrix}$$

- Initial conditions: $x_1 = 1$, all others are 0.





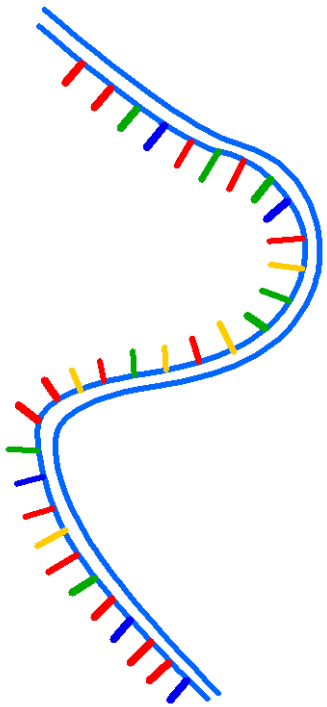


- The most frequent genotype is not necessarily the one with highest A_i.
- The steady state population distribution is called the **quasispecies.**
  - That means, a vertical slice

GARD model (Graded Autocatalysis Replication Domain)

RNA world

Lipid world



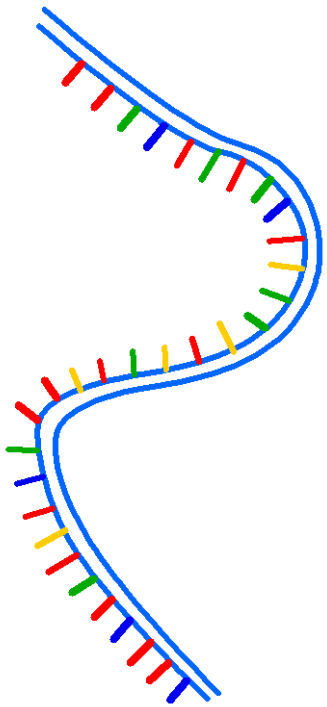
DNA / RNA / Polymers →

<u>Sequence</u>

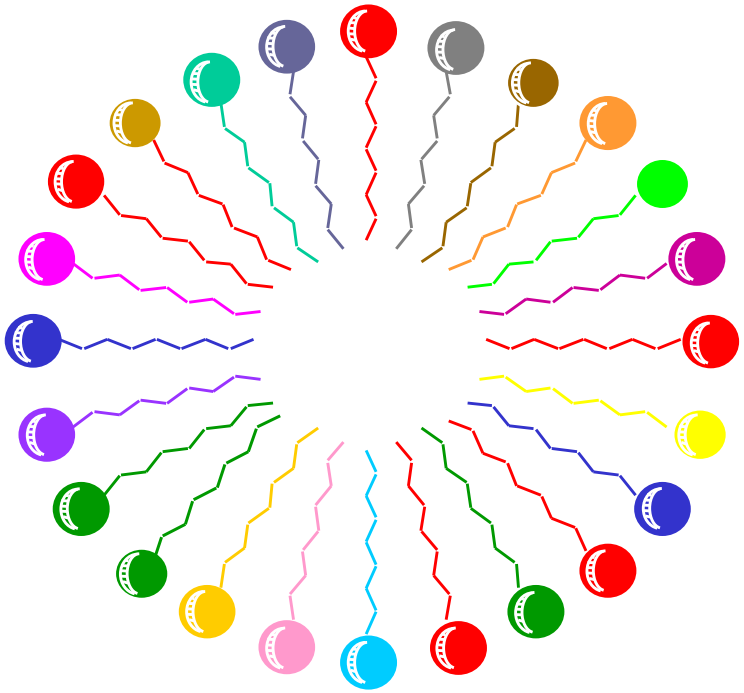
covalent bonds

GARD model (Graded Autocatalysis Replication Domain)

RNA world



Lipid world



DNA / RNA / Polymers →
<u>Sequence</u>
covalent bonds

Assemblies / Clusters / Vesicles / Membranes →
<u>Composition</u>
non-covalent bonds

Segre and Lancet, EMBO Reports 1 (2000)

GARD model (Graded Autocatalysis Replication Domain)

- Synthetic chemistry
- Kinetic model
- Catalytic network (β) of rate-enhancement values

Segre, Ben-Eli and Lancet, Proc. Natl. Acad. Sci. 97 (2000)

GARD model (Graded Autocatalysis Replication Domain)

- Synthetic chemistry
- Kinetic model
- Catalytic network (β) of rate-enhancement values

Rate enhancement

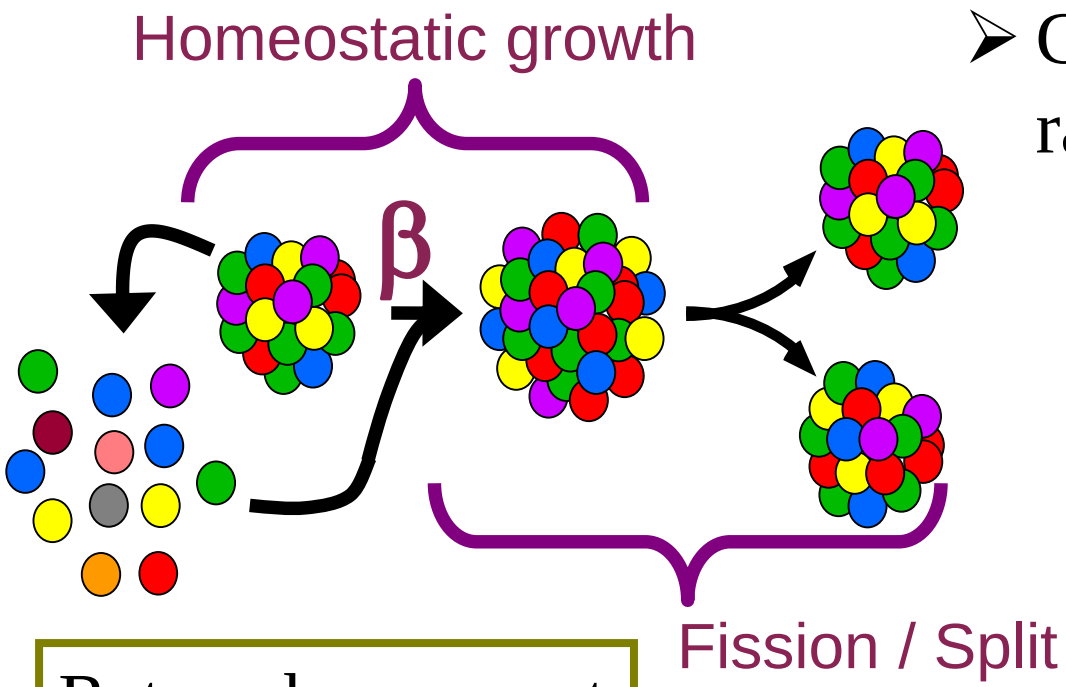
$$\frac{dn_i}{dt} = \left(k_f \rho_i N - k_b n_i\right)\left(1 + \sum_{j=1}^{N_G} \beta_{ij} \frac{n_j}{N}\right) \qquad (i = 1..N_G)$$

Molecular repertoire

Segre, Ben-Eli and Lancet, Proc. Natl. Acad. Sci. 97 (2000)

GARD model (Graded Autocatalysis Replication Domain)

- Synthetic chemistry
- Kinetic model
- Catalytic network (β) of rate-enhancement values



Rate enhancement

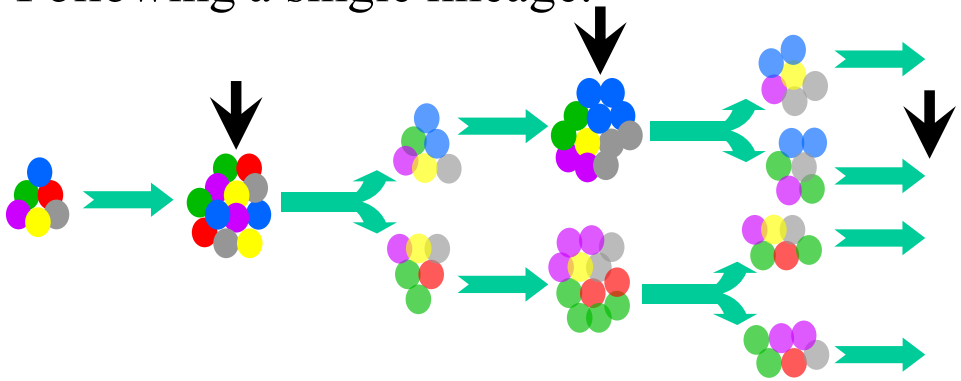
$$\frac{dn_i}{dt} = \left(k_f \rho_i N - k_b n_i\right)\left(1 + \sum_{j=1}^{N_G} \beta_{ij} \frac{n_j}{N}\right) \qquad (i = 1..N_G)$$

Molecular repertoire

Segre, Ben-Eli and Lancet, Proc. Natl. Acad. Sci. 97 (2000)

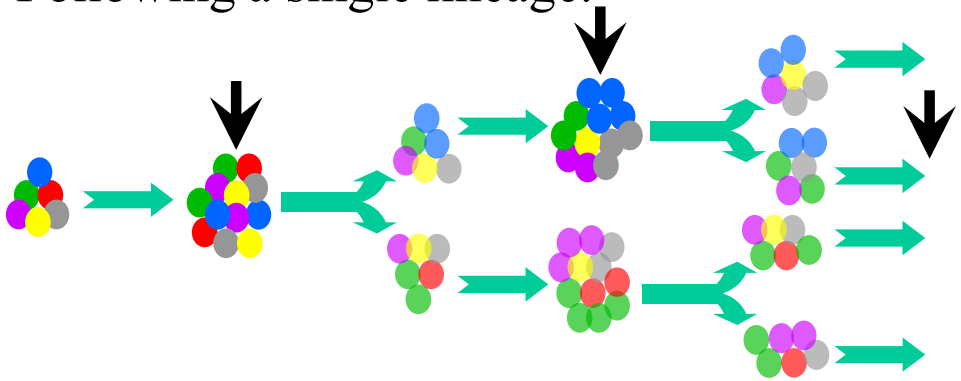
GARD model (Graded Autocatalysis Replication Domain)

Following a single lineage.

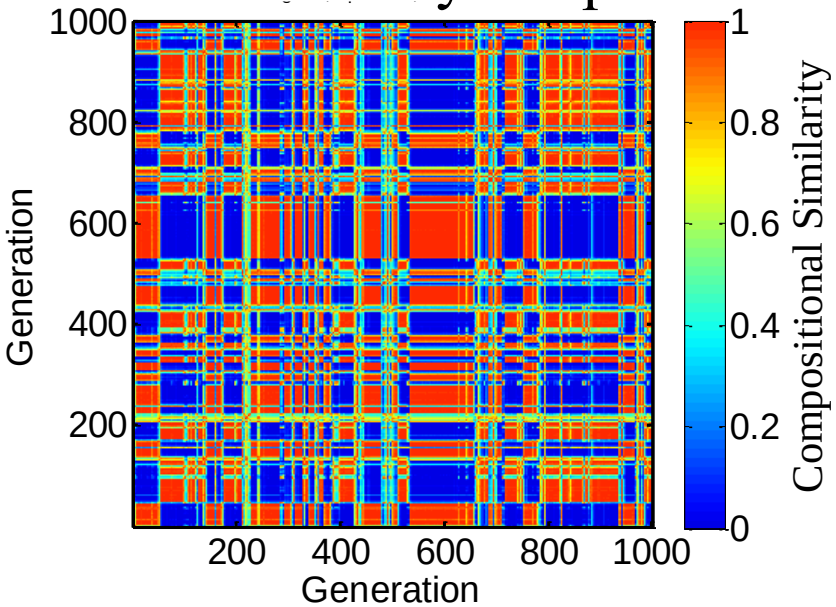


GARD model (Graded Autocatalysis Replication Domain)

Following a single lineage.

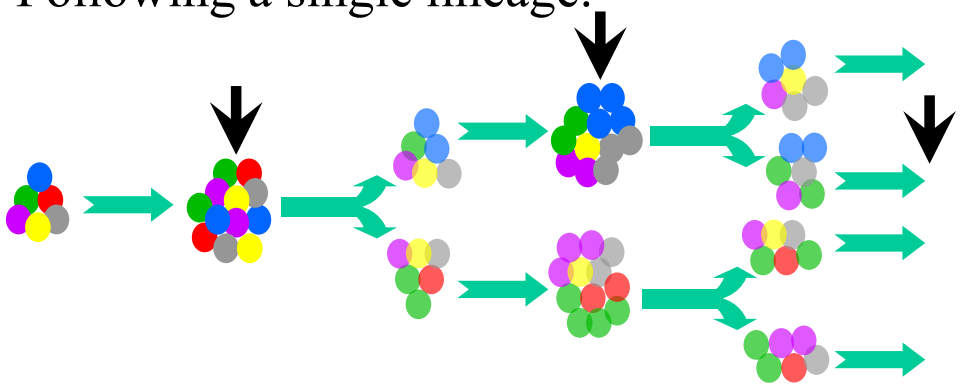


Similarity 'carpet'

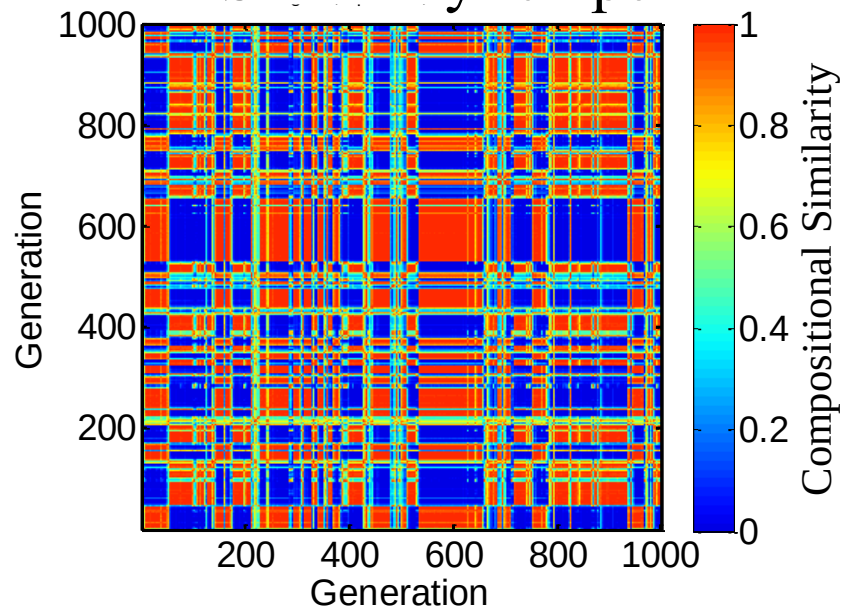


GARD model (Graded Autocatalysis Replication Domain)

Following a single lineage.



Similarity 'carpet'



Composome (compositional genome): a faithfully replicating composition/assembly.

Compotype (composome type): a collection of similar composomes.

Molecular Compotype: the center of mass of the compotype cloud, treated as a molecular assembly.

GARD model (Graded Autocatalysis Replication Domain)

$$\frac{dn_i}{dt} = \left(k_f \rho_i N - k_b n_i\right)\left(1 + \sum_{j=1}^{N_G} \beta_{ij} \frac{n_j}{N}\right) \qquad \left(i = 1 \ldots N_G\right)$$

- Idea: if we ignore the stochasticity inherent in the model, then errorless-replication occurs according to beta matrix eigenvectors.

Moment Algebraux

- Multiplying a matrix by a vector gives another vector
- $\boldsymbol{A} \cdot \vec{x} = \vec{y}$
- $\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{21}x_1 + a_{22}x_2 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$

- An **eigenvector** is a vector $\vec{x}$ such that:
- $\boldsymbol{A} \cdot \vec{x} = \lambda \vec{x}$
  - λ is called the **eigenvalue**.
  - λ may be complex, and so may the eigenvector.

Moment Algebraux

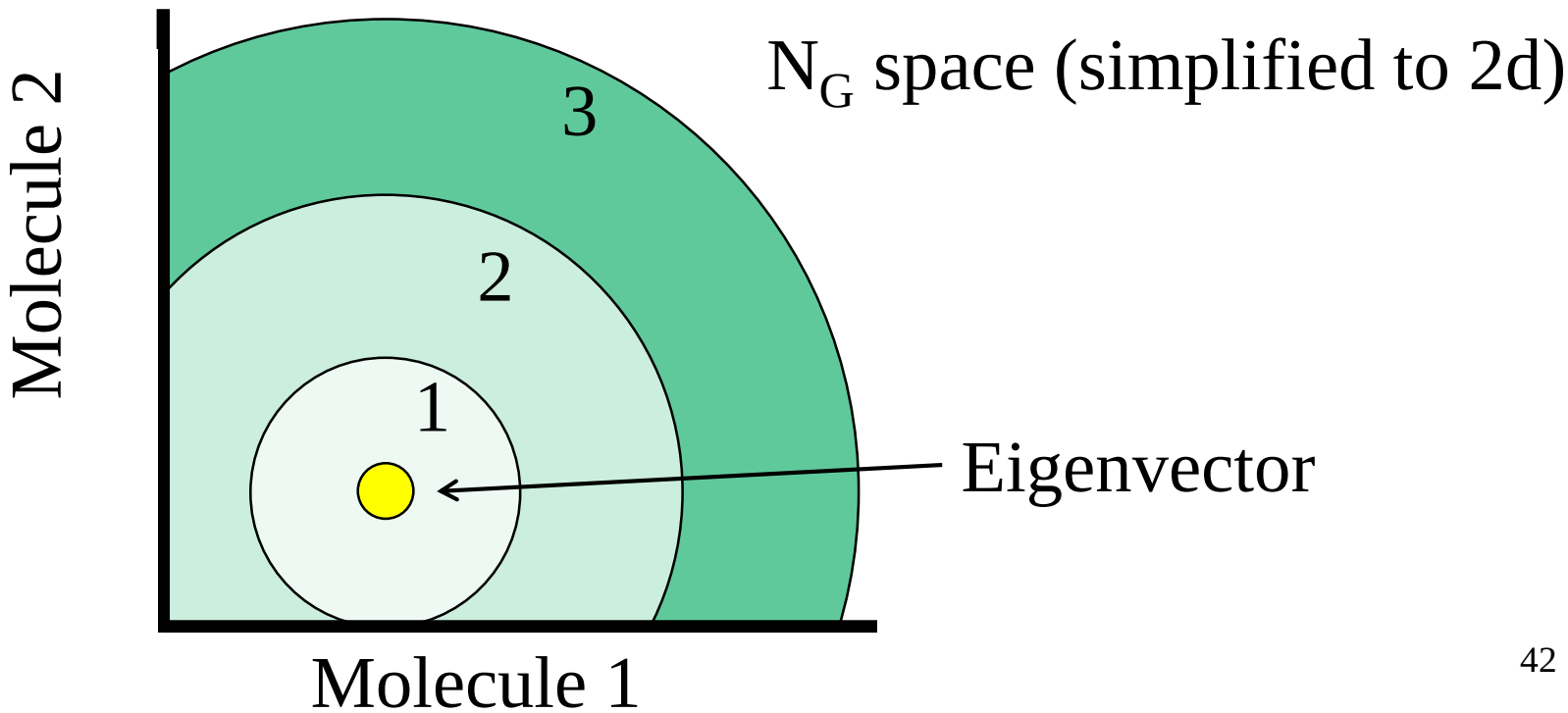
- The Perron-Frobenius Theorem:
  - A matrix with strictly positive entries contains a maximal **real** eigenvalue.
  - Its eigenvector is real and non-negative. In fact, it's the only one with this property.
- As molecular assemblies must contain real non-negative number of molecules, this looks interesting.

GARD and Quasispecies

- Do GARD population dynamics behave like the quasispecies model?
- What we would like to do:
  - For each assembly, experimentally find out the transition frequencies and replication rates
  - In other words: find **Q** and **A.**
- Problem:
  - $N_G = 100$, $n_{max} = 100$ ➔ There are $\binom{199}{100}$ possible assemblies ($\sim 4 \cdot 10^{58}$)

GARD and Quasispecies

- Solution: group some assemblies together and treat them as one genotype.
- We decided to group together by distances from the eigenvector.

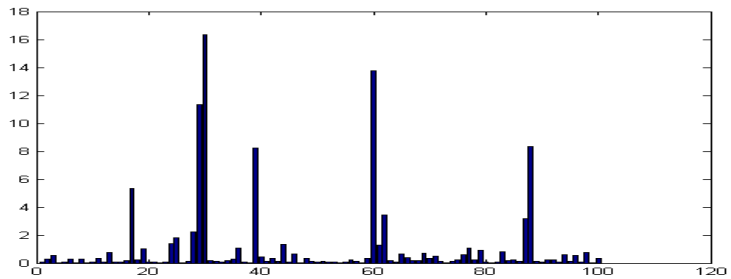


GARD and Quasispecies

- Problem: how do we sample such a large space?
- 30000 assemblies are randomly generated.

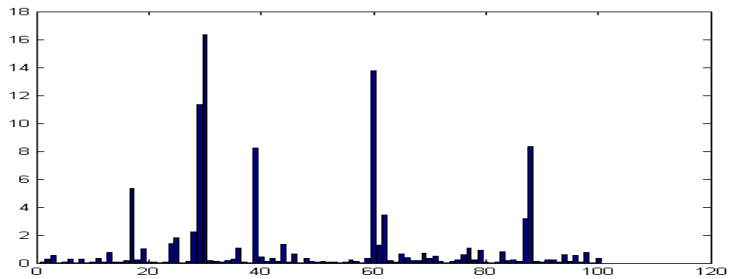
GARD and Quasispecies

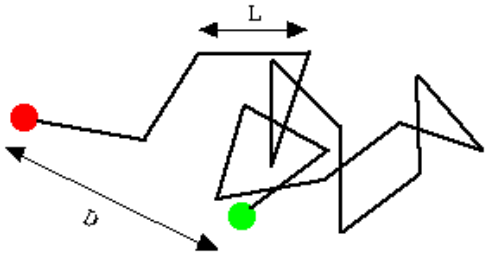
- Problem: how do we sample such a large space?
- 30000 assemblies are randomly generated.
  - By filling up assemblies until they reach n_{max}.
    - Assemblies generated this way are **far** from the target.



GARD and Quasispecies

- Problem: how do we sample such a large space?
- 30000 assemblies are randomly generated.
  - By filling up assemblies until they reach n_{max}.
    - Assemblies generated this way are **far** from the target.
  - By starting at eigenvector and random walking.
    - Assemblies generated this way are **close** to the target.

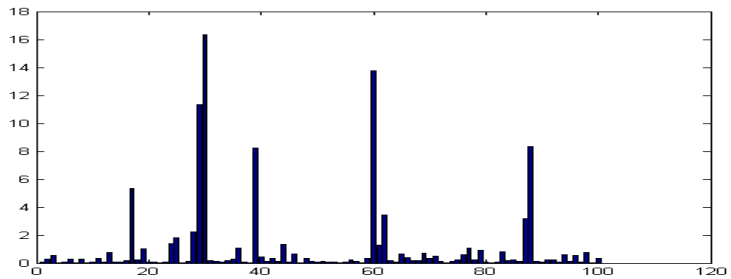


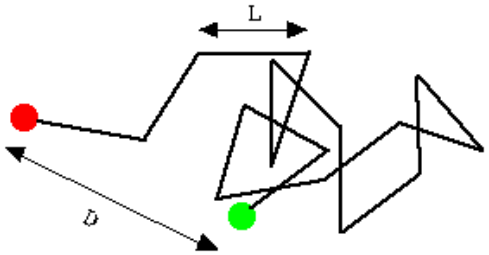


- http://www2.ess.ucla.edu/~jewitt/oort2-random.html

GARD and Quasispecies

- Problem: how do we sample such a large space?
- 30000 assemblies are randomly generated.
  - By filling up assemblies until they reach n_{max}.
    - Assemblies generated this way are **far** from the target.
  - By starting at eigenvector and random walking.
    - Assemblies generated this way are **close** to the target.



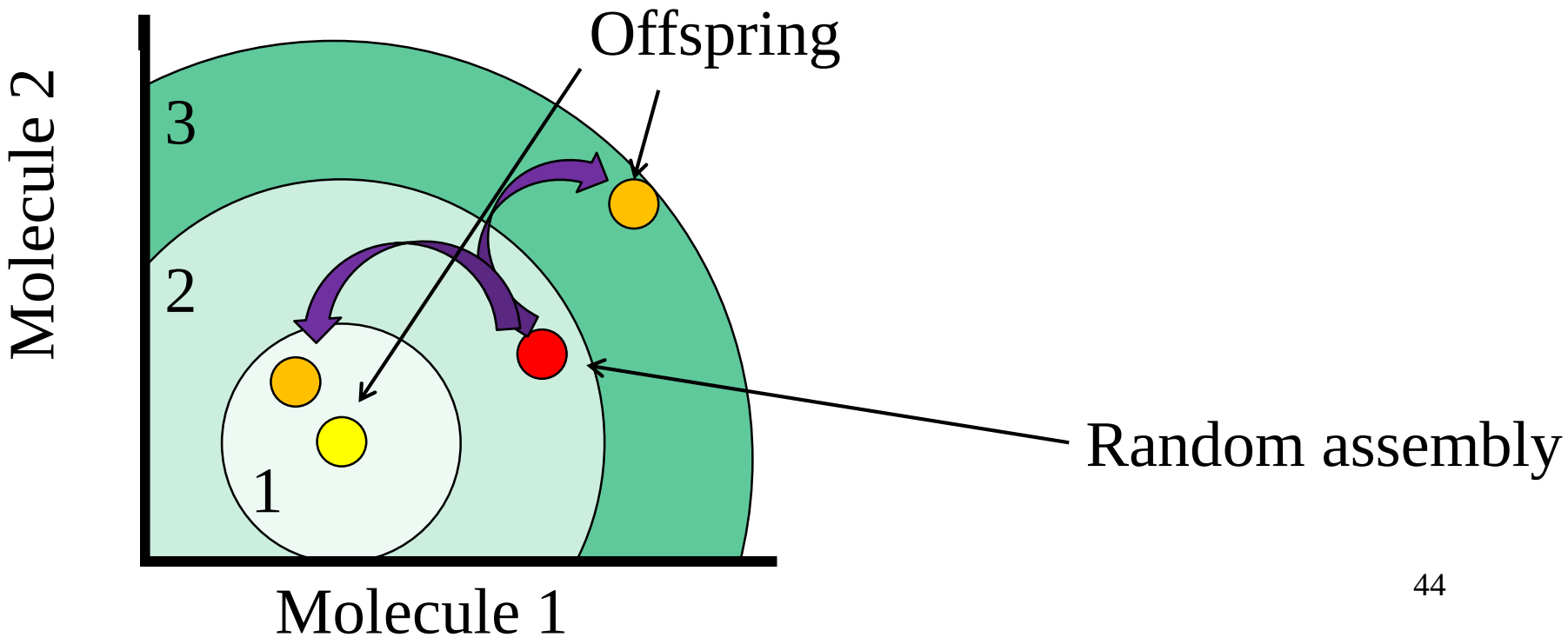


- Sampling is still a problematic issue.

- http://www2.ess.ucla.edu/~jewitt/oort2-random.html

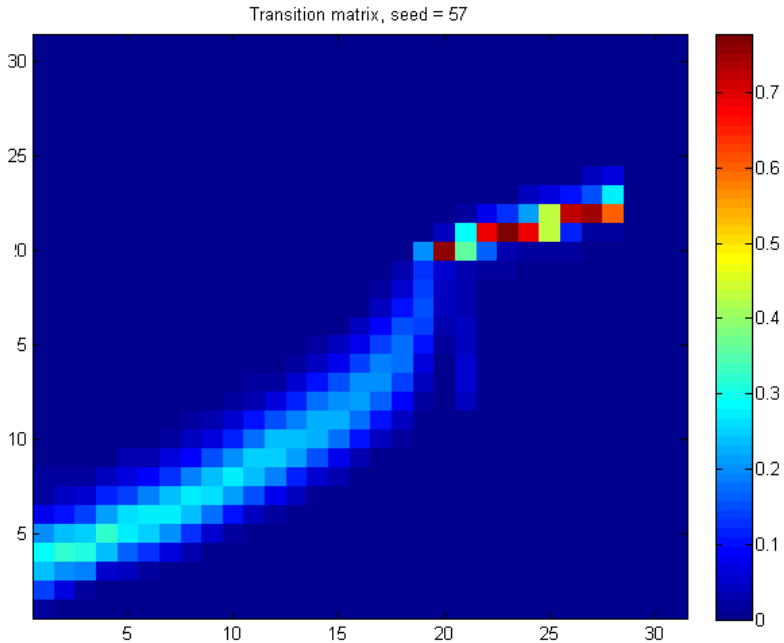
GARD and Quasispecies

- Each assembly is split and its offspring grown.
  - **Q** = to where did the assembly split?
  - **A** = how long did it take the offspring to grow?



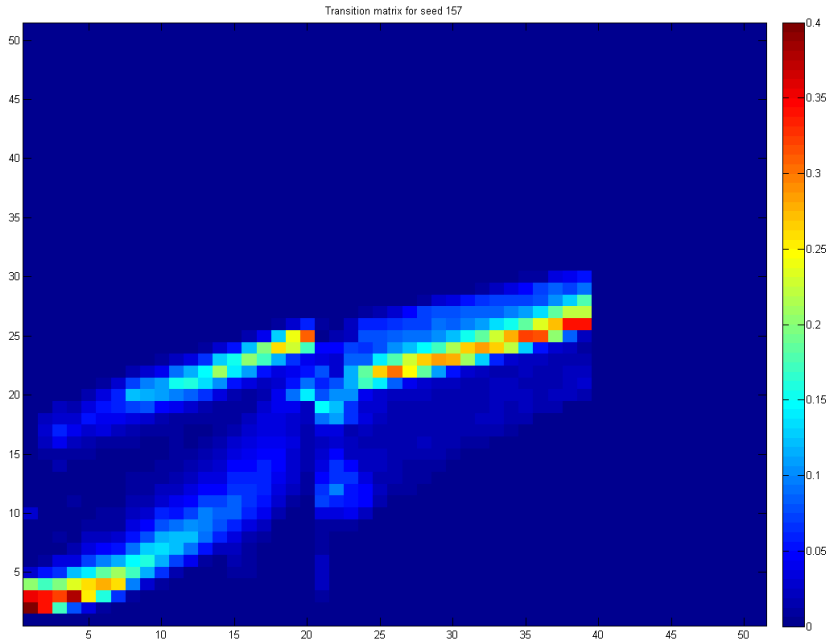
- Example of a transition matrix:

To



From

To



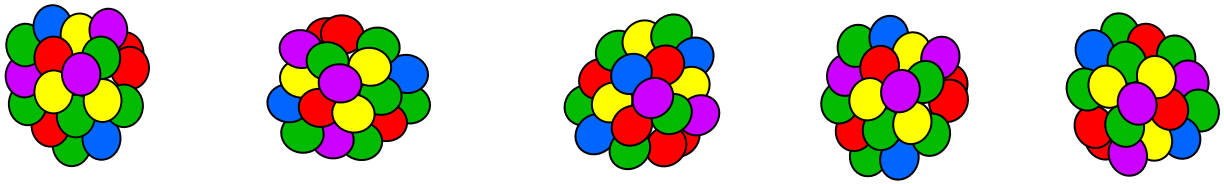
From

GARD and Quasispecies

- Now all that is left is to compare population model with quasispecies shell model.
- The population runs were already performed by Omer: a constant population Moran-process.

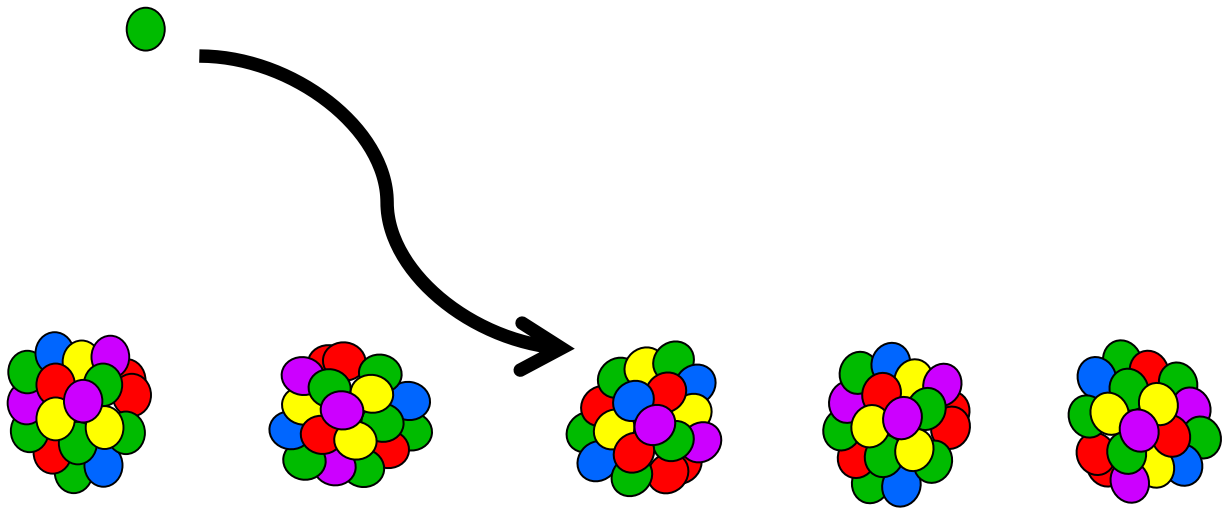
GARD and Quasispecies

- There is a constant number of assemblies in a population.



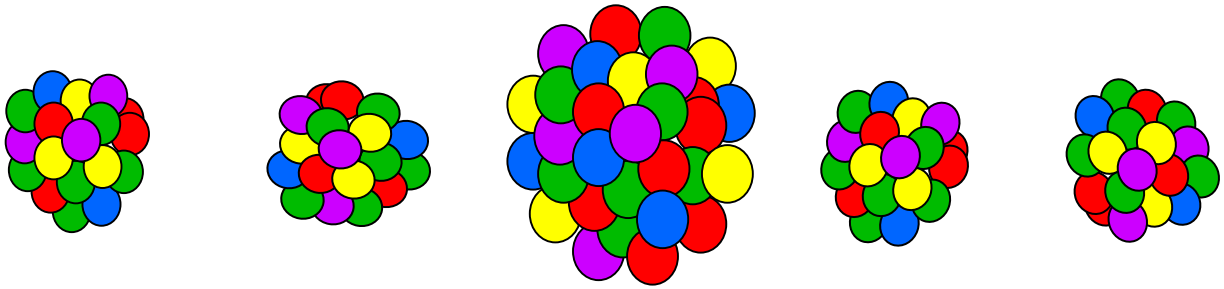
GARD and Quasispecies

- There is a constant number of assemblies in a population.
- Each turn, a single molecule is added.



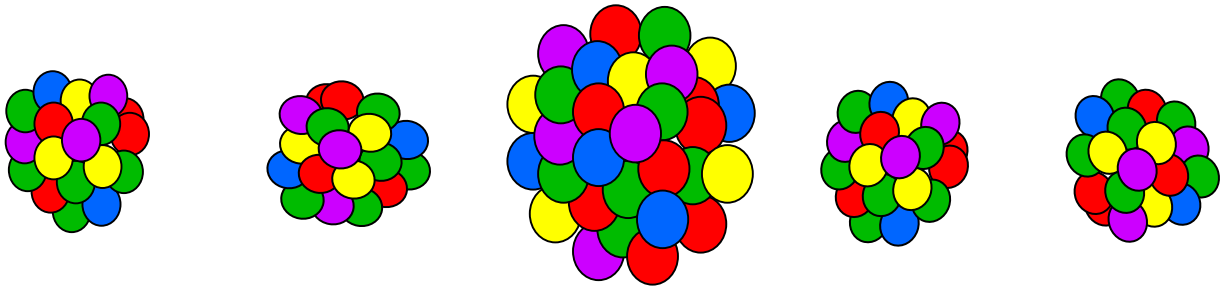
GARD and Quasispecies

- There is a constant number of assemblies in a population.
- Each turn, a single molecule is added.



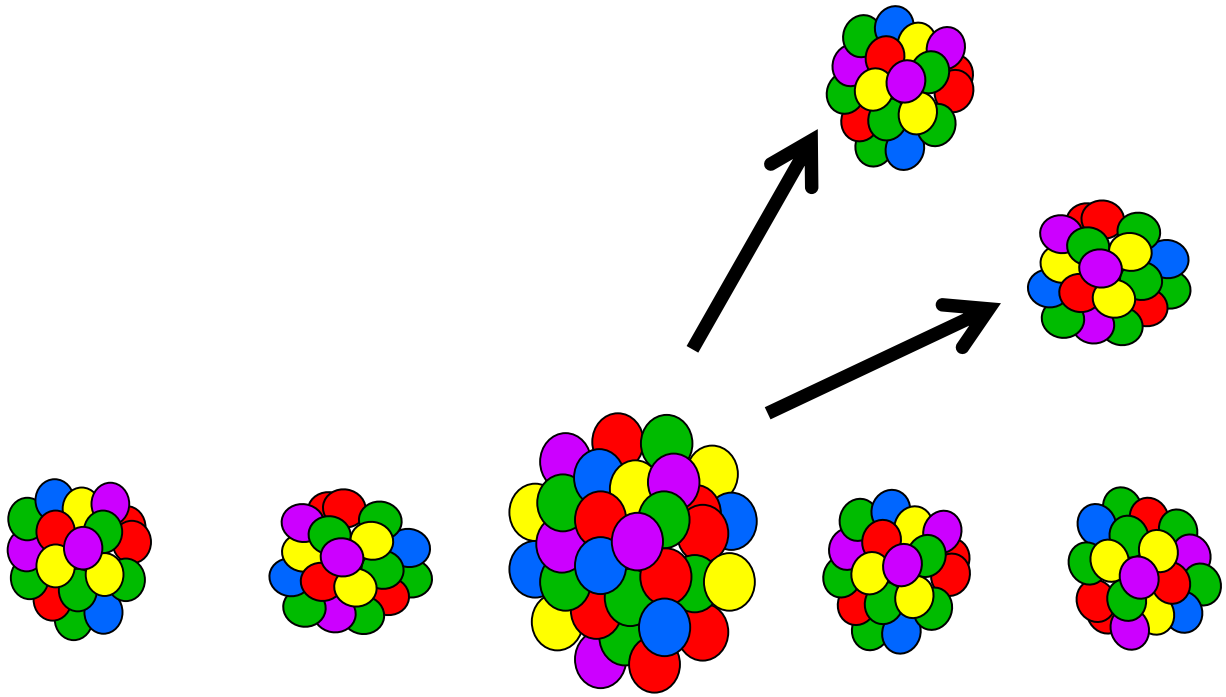
GARD and Quasispecies

- There is a constant number of assemblies in a population.
- Each turn, a single molecule is added.
- Upon split, the parent as well as a random assembly are replaced with the two children.



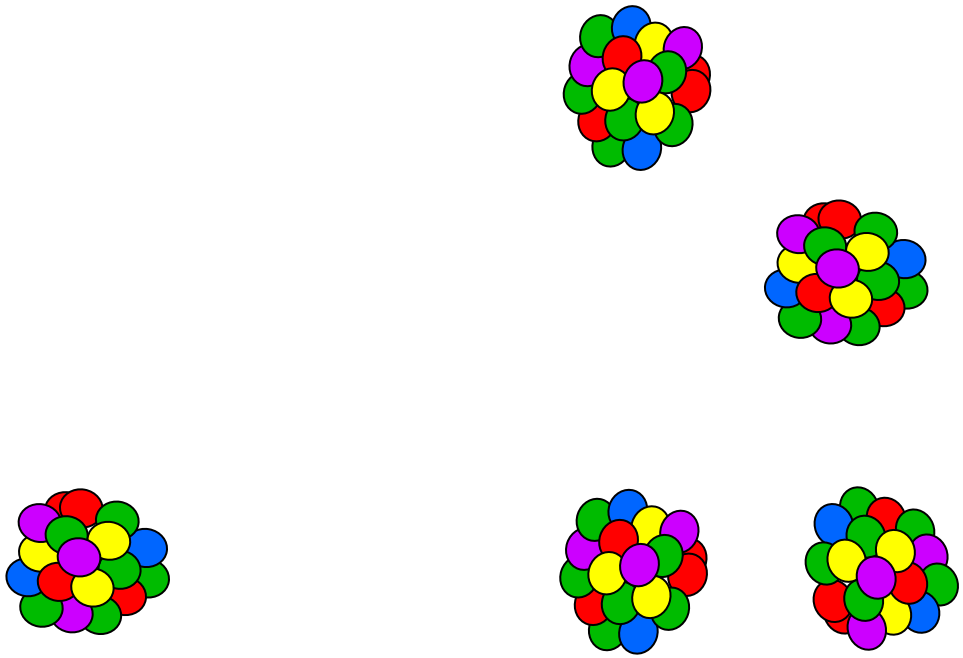
GARD and Quasispecies

- There is a constant number of assemblies in a population.
- Each turn, a single molecule is added.
- Upon split, the parent as well as a random assembly are replaced with the two children.



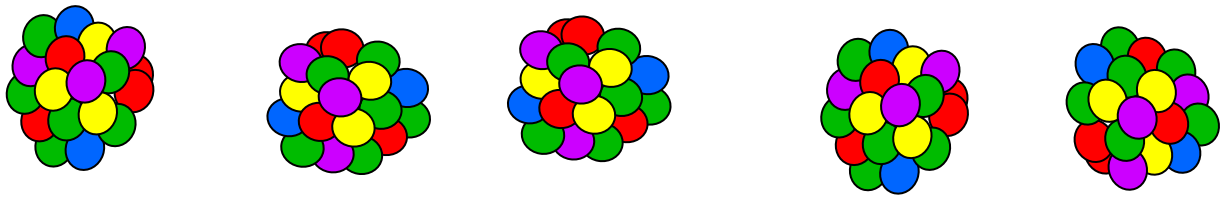
GARD and Quasispecies

- There is a constant number of assemblies in a population.
- Each turn, a single molecule is added.
- Upon split, the parent as well as a random assembly are replaced with the two children.



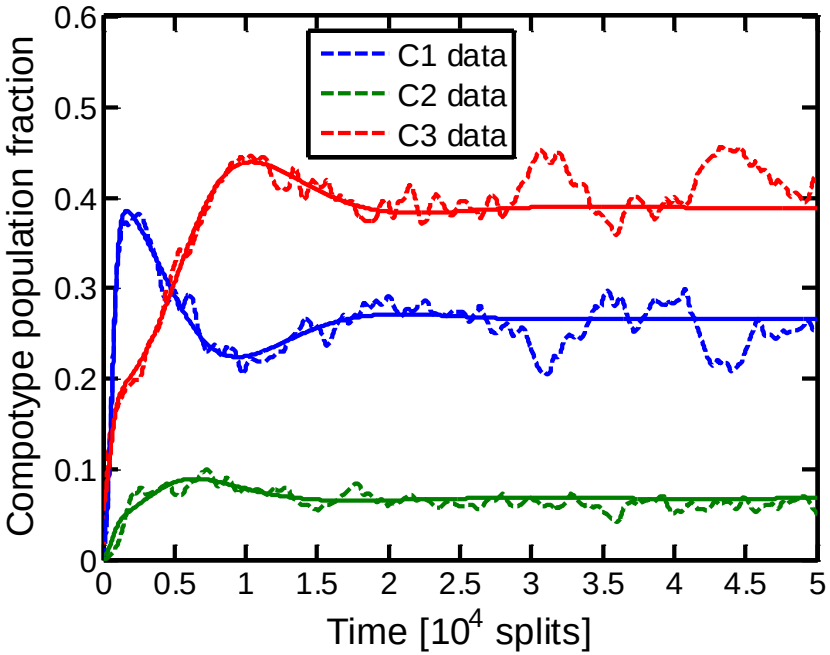
GARD and Quasispecies

- There is a constant number of assemblies in a population.
- Each turn, a single molecule is added.
- Upon split, the parent as well as a random assembly are replaced with the two children.



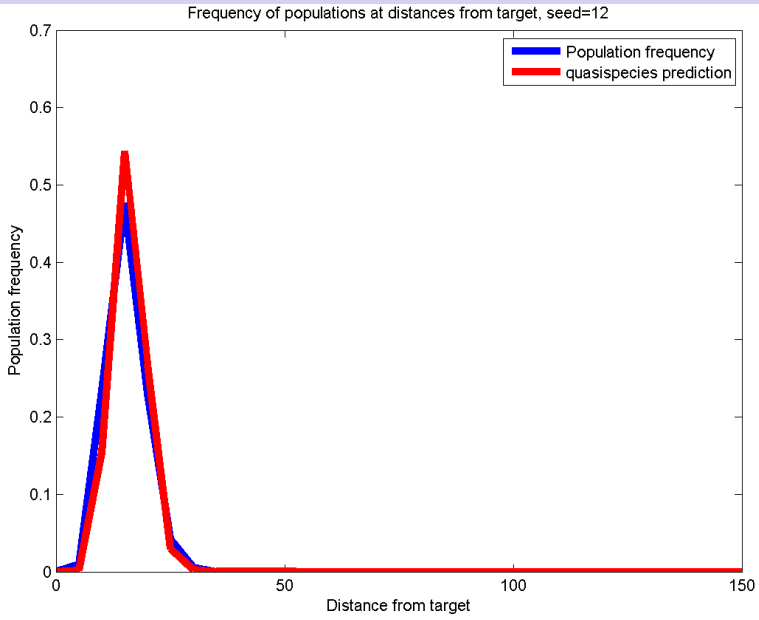
GARD and Quasispecies

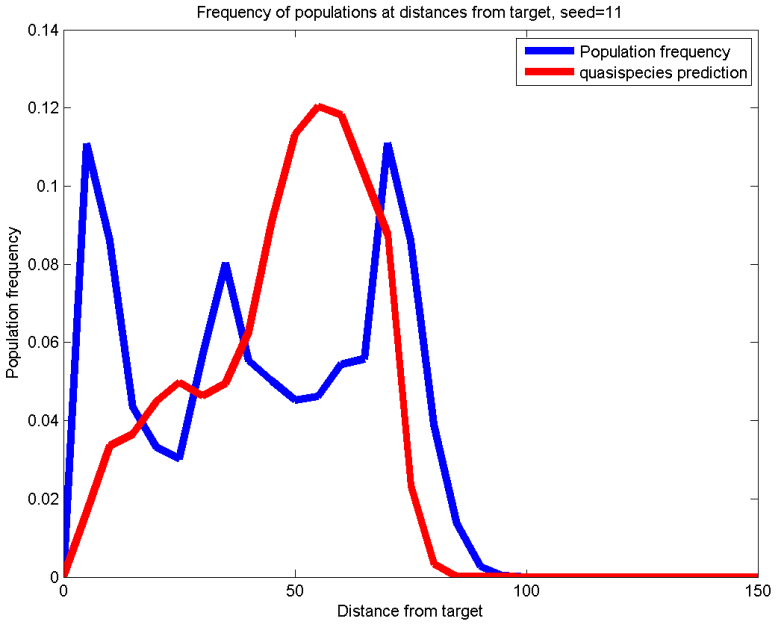
- The population simulation goes into steady state, concerning the frequencies of **compotypes** (as shown by Omer)

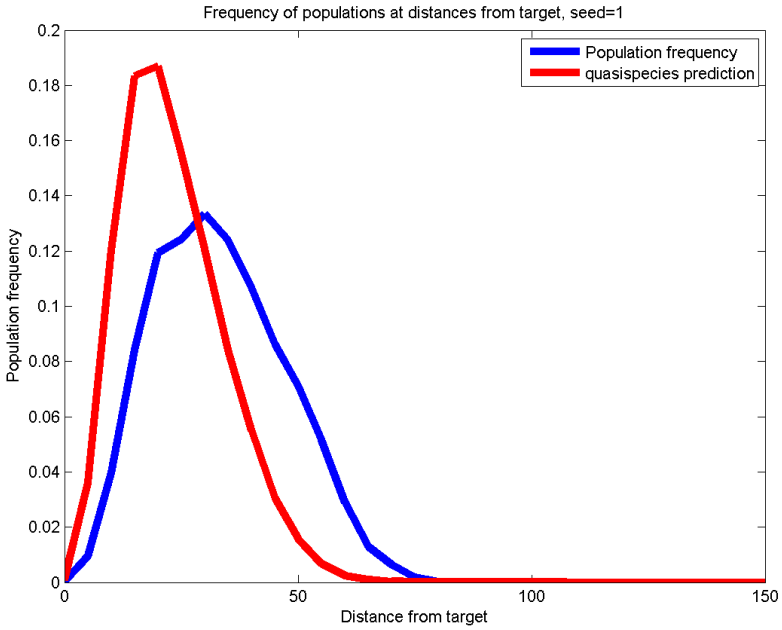


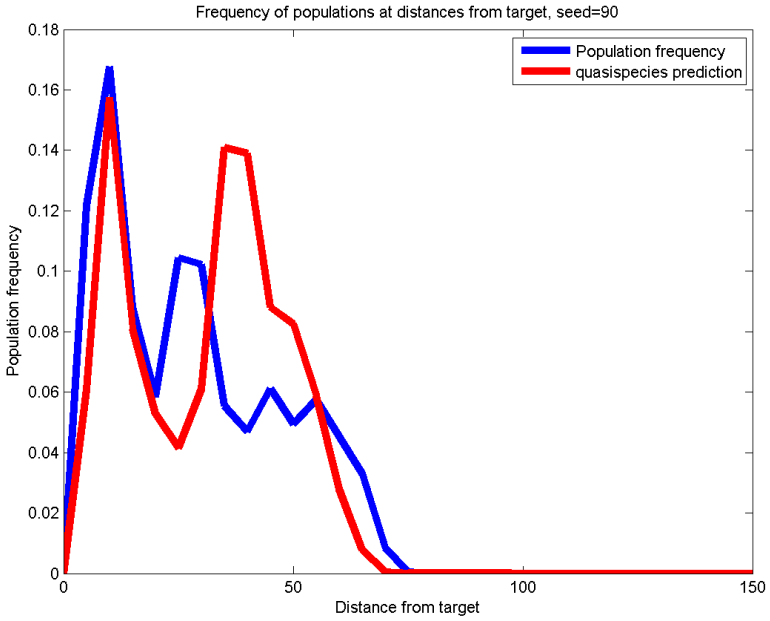
- The distribution of distances from the eigenvector was calculated for the population steady state, and compared with prediction.

GARD and Quasispecies









GARD and Quasispecies

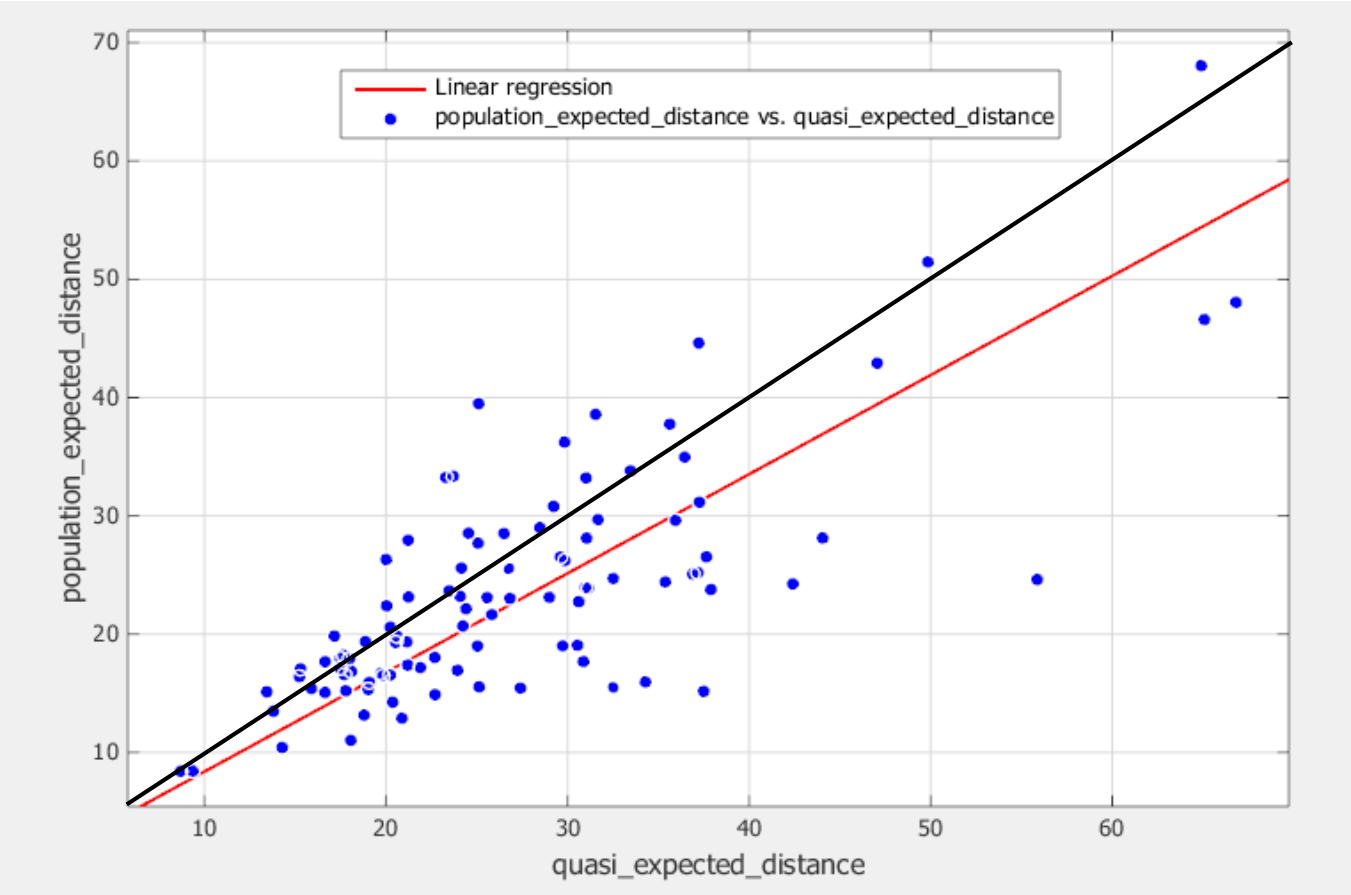
- How do we know if we got a good match?
- Two metrics were used:
  - Expected distance:

$$E = \sum_{i=1}^{n} d_i x_i$$

  - Pearson correlation – do the troughs and hills go up and down at the same time for both population and prediction?

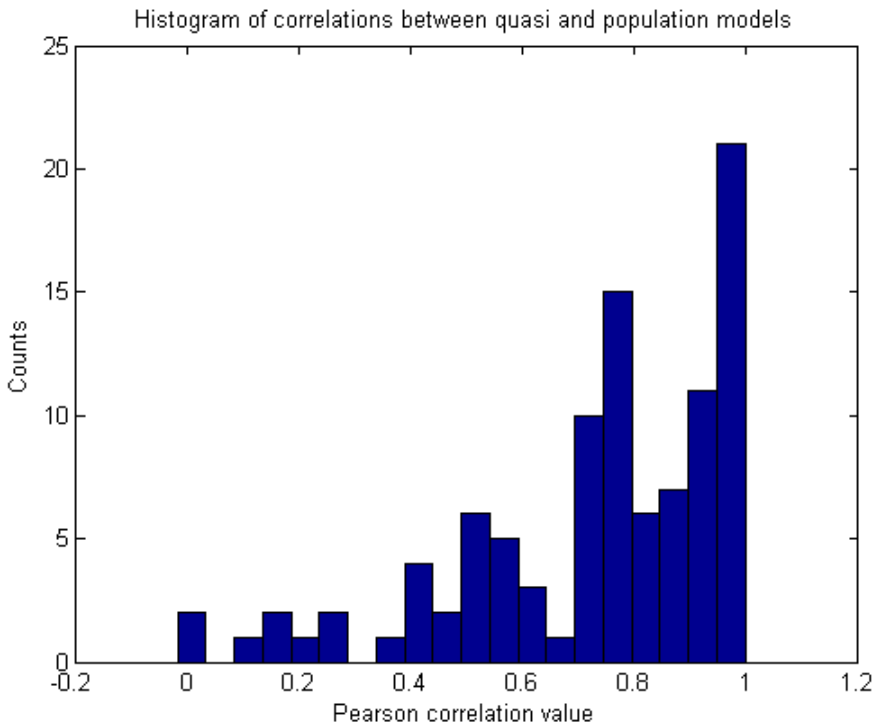
GARD and Quasispecies

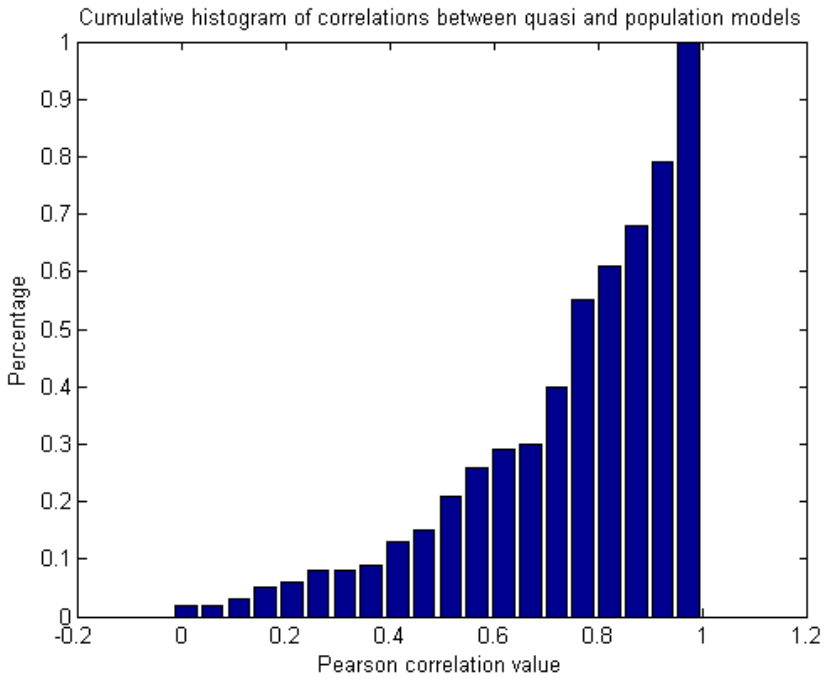
- Expected distances
- $R^2 = 0.52$, slope = 0.83



GARD and Quasispecies

- Correlations
- About half are above 0.8





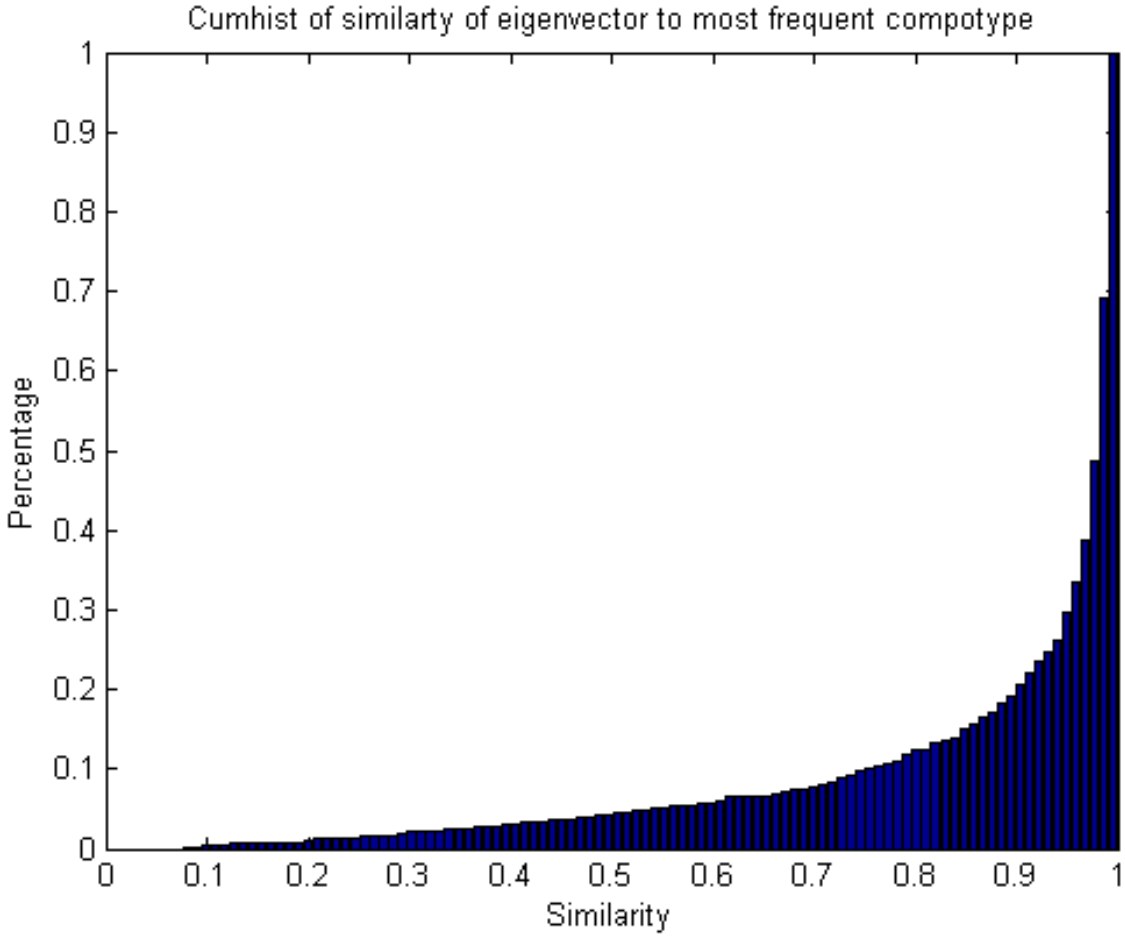
GARD and Quasispecies

- Ok, is this good?
- We can change the target of the distance measurement, to see if we get a better result.

- Two more assemblies were tried:
  - The most common compotype
  - A random assembly.

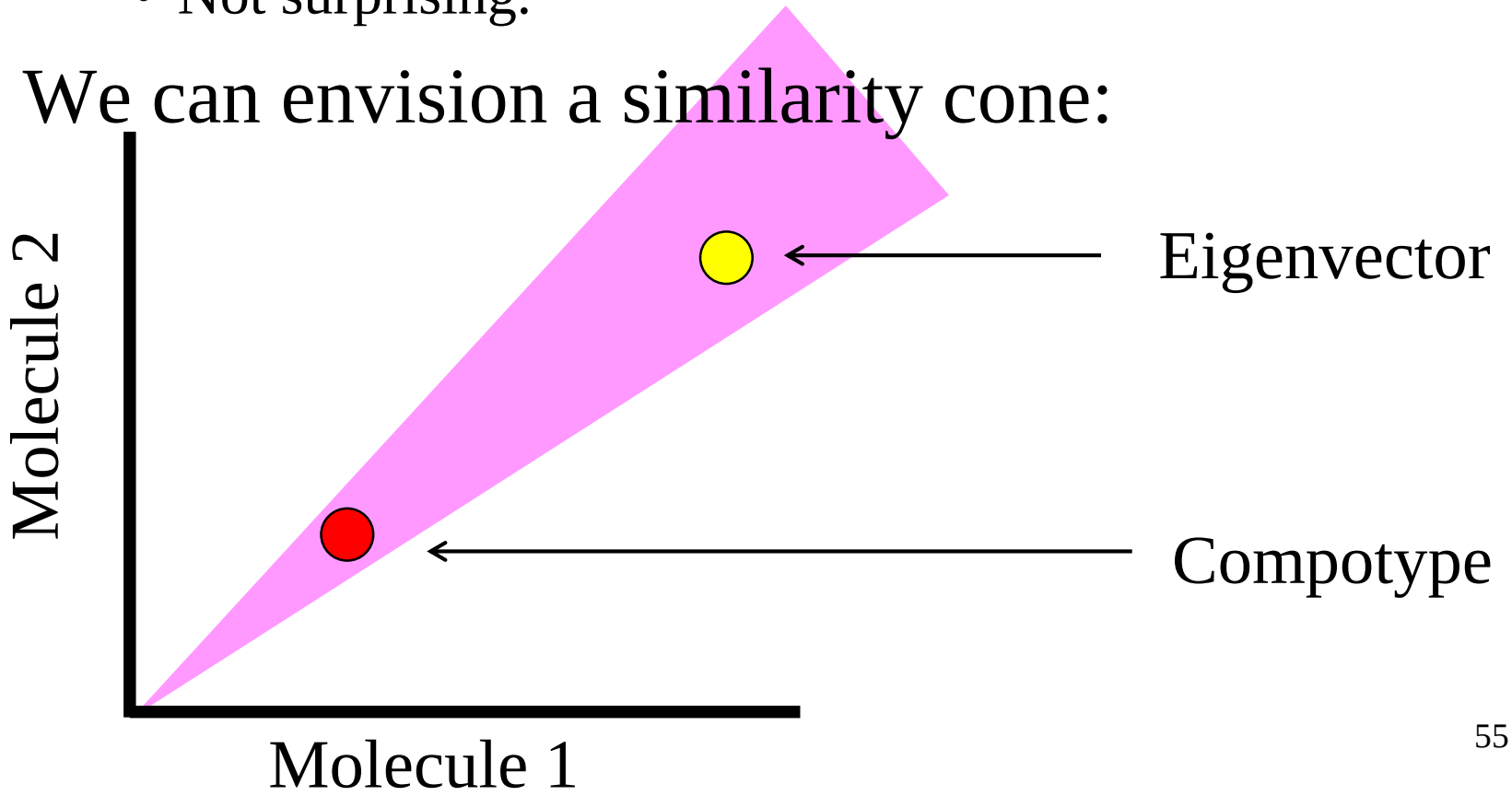
GARD and Quasispecies

- The most common compotype is very similar to the eigenvector.



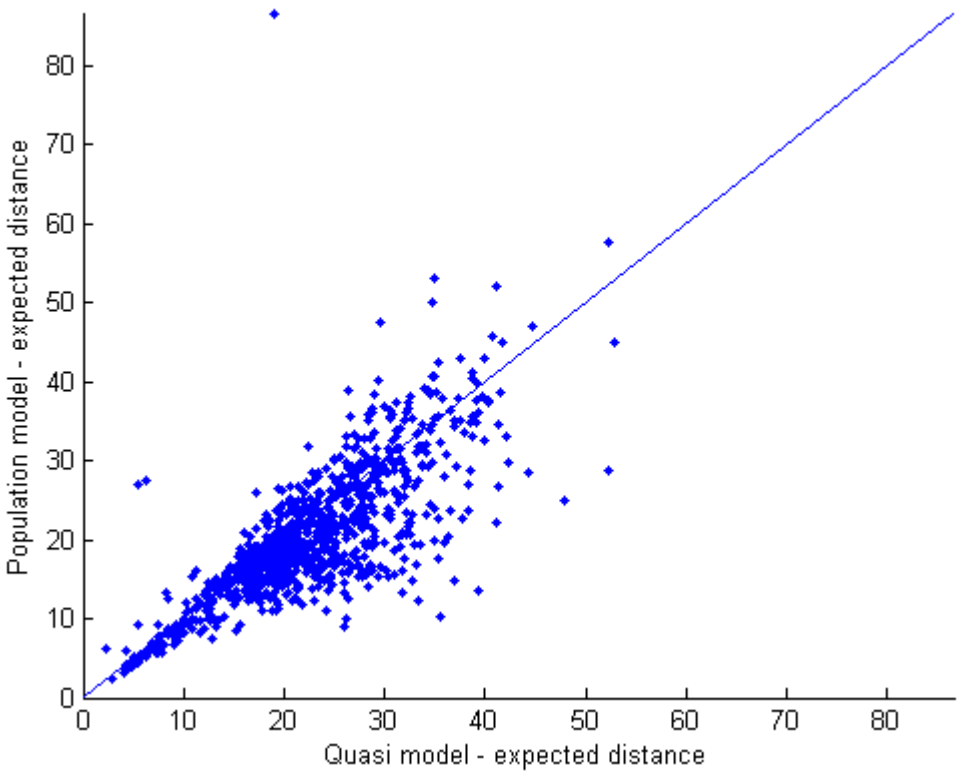
GARD and Quasispecies

- However, they are not exactly the same; often the eigenvector is larger (larger Euclidean norm)
  - This means it is less homogenous than compotypes
    - Not surprising.
- We can envision a similarity cone:



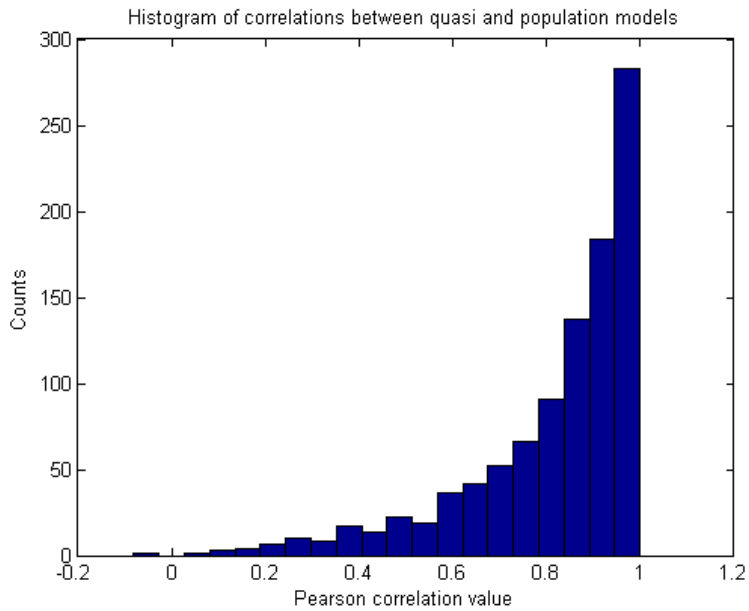
GARD and Quasispecies

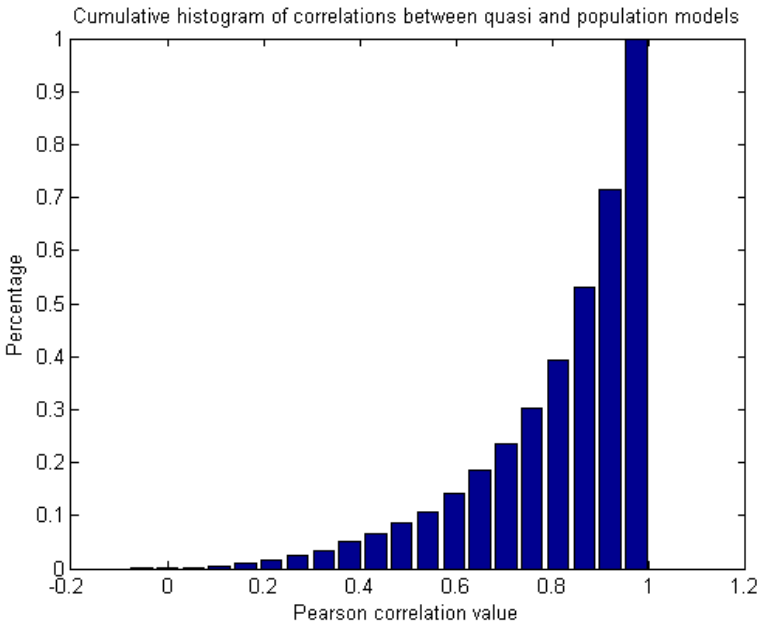
- Results are better for compotypes, and worse for random. $R^2 = 0.64$, slope = 0.89



GARD and Quasispecies

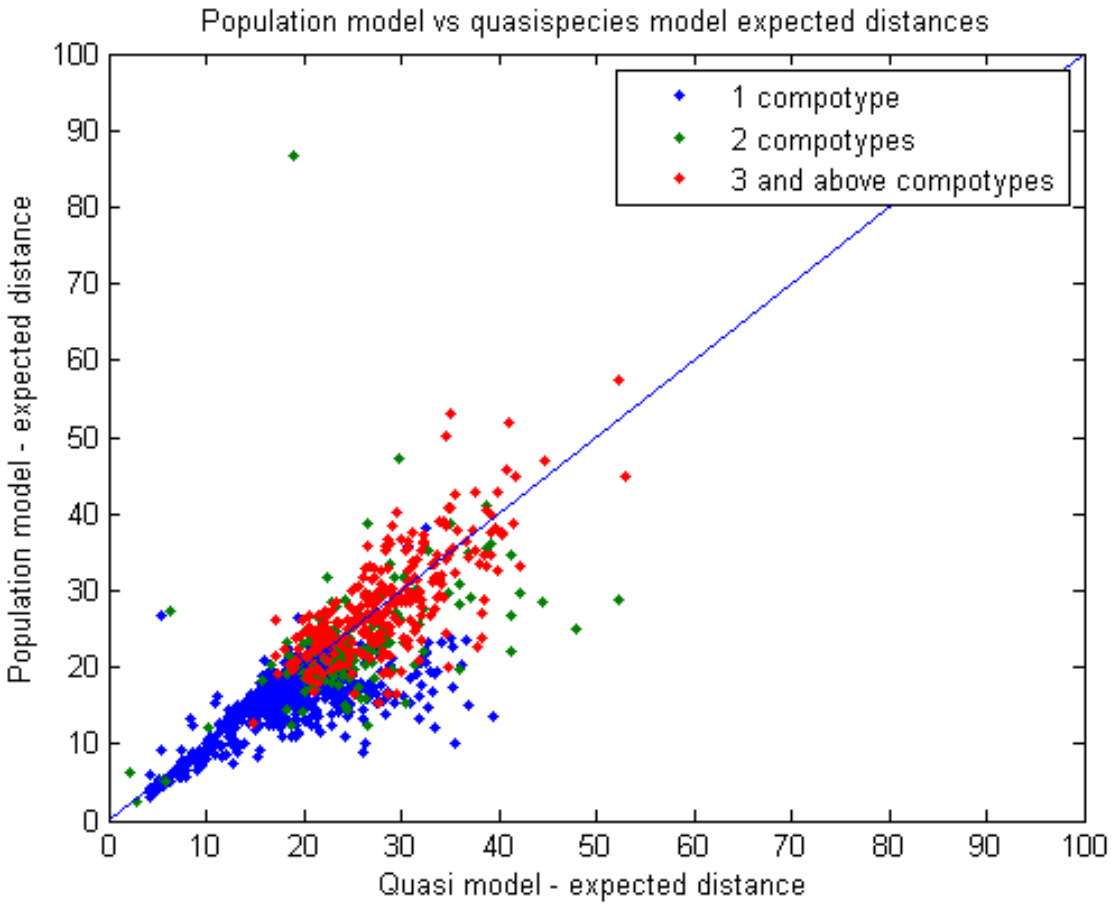
- Correlations
- About 0.7 are above 0.8





GARD and Quasispecies

- There is a dependence on the number of compotypes

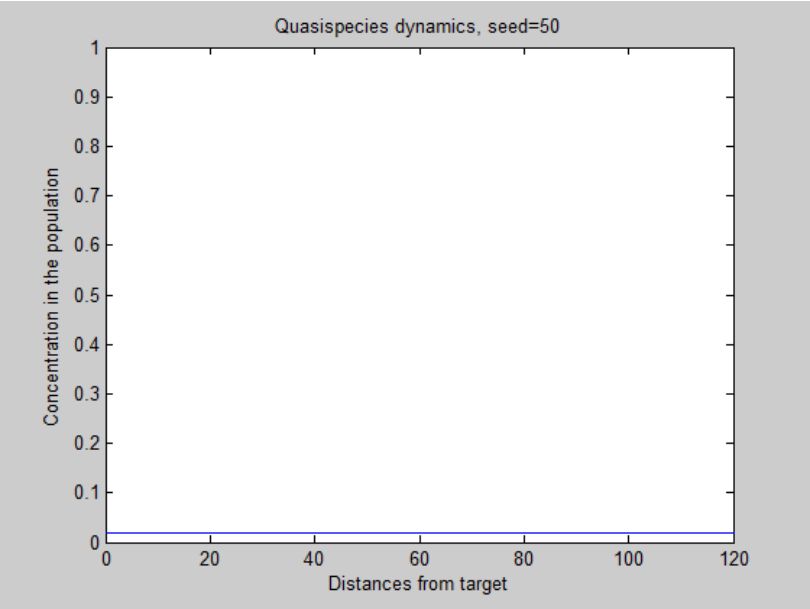


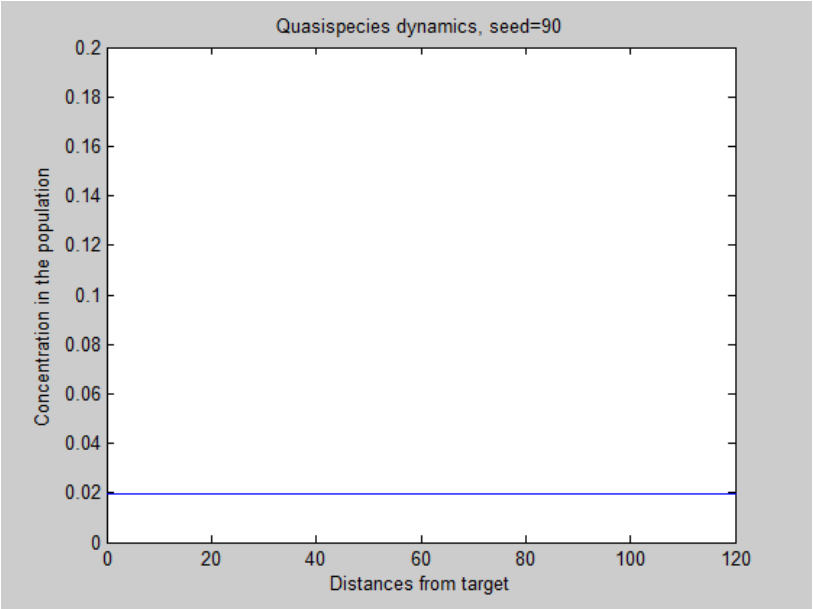
GARD and Quasispecies

- The future…?
  - Better sampling
  - More rigorous analysis of number of compotypes

GARD and Quasispecies

- The future…?
  - Better sampling
  - More rigorous analysis of number of compotypes
  - Dynamics, and not just steady state





Error Catastrophe in GARD

- For sequential information carriers, *q* acts as a "faithful replication" parameter.
- Does anything like this exist for GARD?

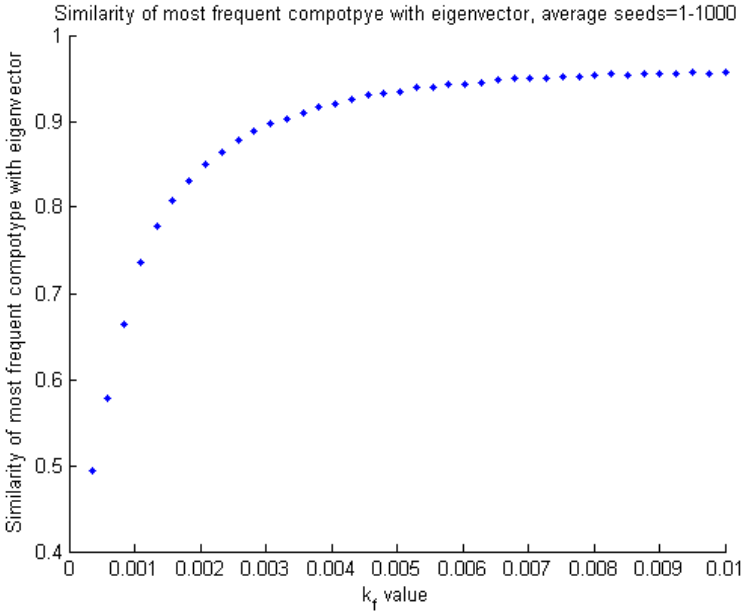
  "*Idea: if we ignore the stochasticity inherent in the model and solution, then errorless-replication occurs according to beta matrix eigenvectors.*"

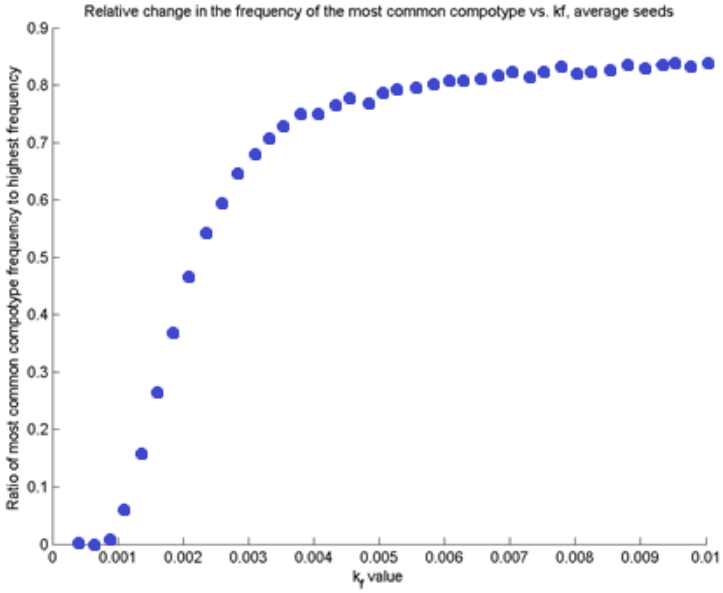
  - R.G

- Forward and backward accretion (k_f and k_b) are responsible for much of the stochasticity.

Error Catastrophe in GARD

- What happens if you lower k_f ?
- Run single lineage simulation with different k_f values.





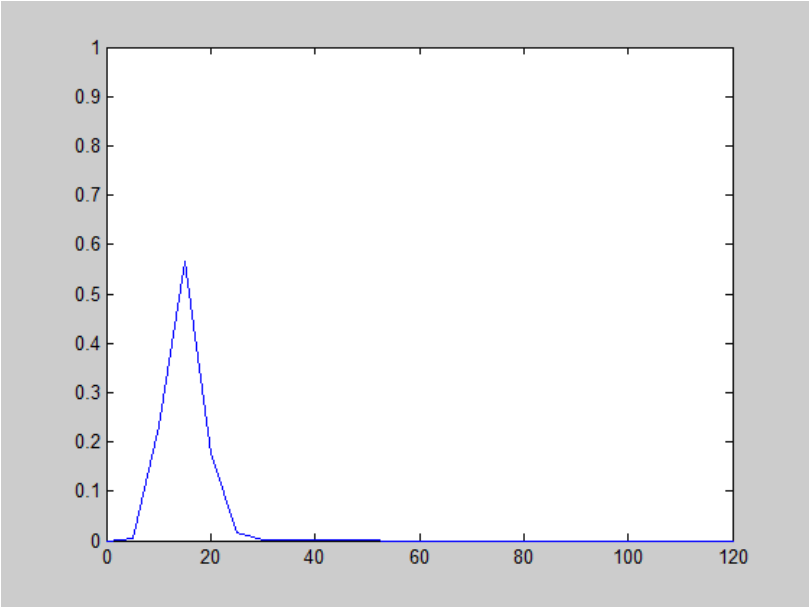
Error Catastrophe in GARD

- Since the most common compotype frequency decreases drastically, there is a high increase in drift → no composomes.

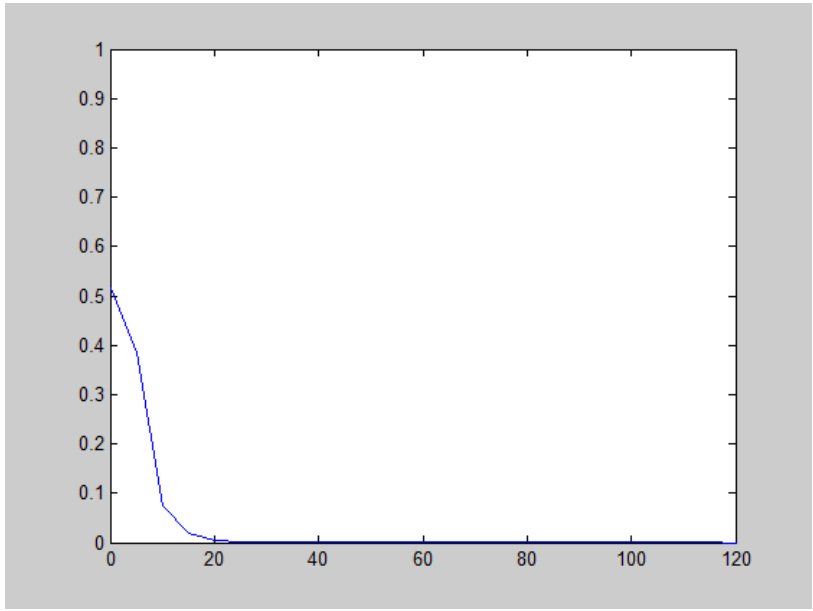
- Conclusion: k_f and k_b affect replication fidelity.

Error Catastrophe in GARD

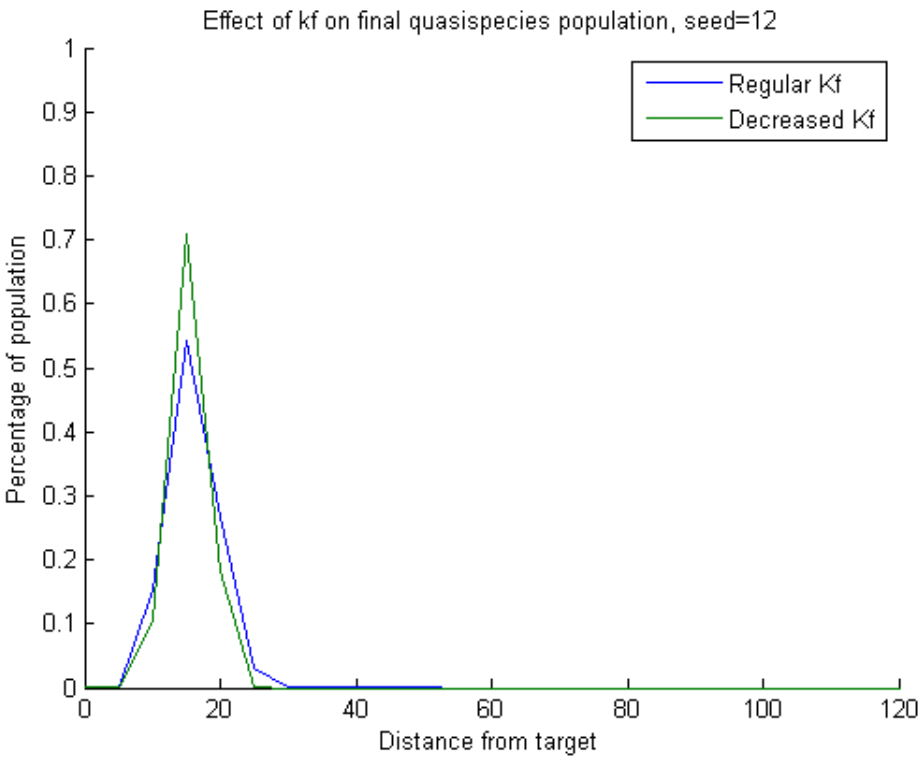
- What happens in quasispecies model?
  - We obtained **Q** and **A** for lower k_f values.
- Two types of results:
- Seed = 12

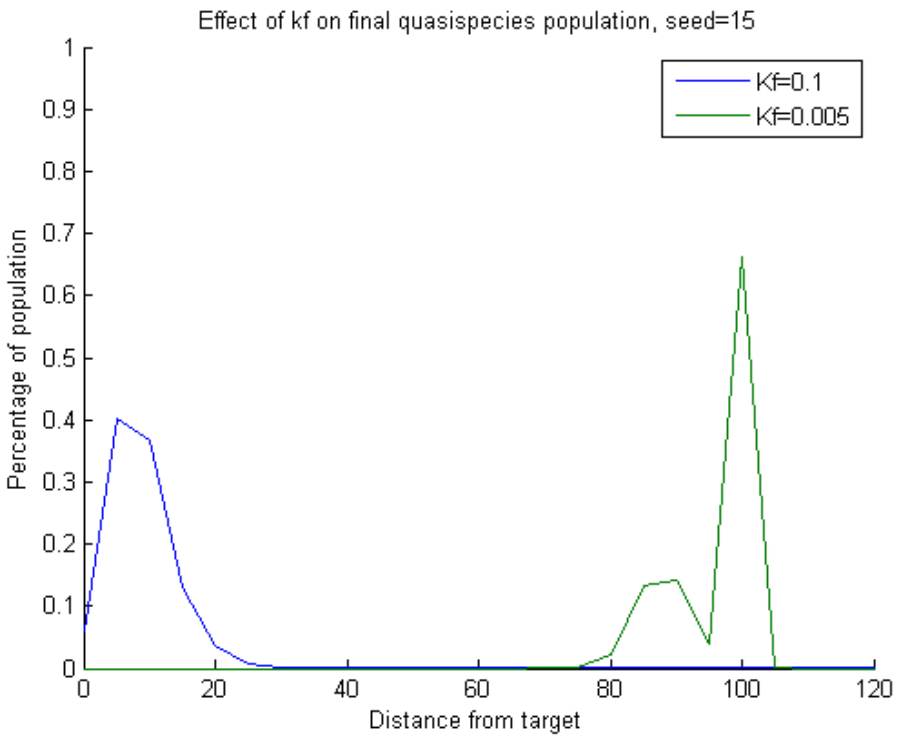


Seed = 15



Error Catastrophe in GARD





- The difference seems to relate to the **size** of the compotype.

Error Catastrophe in GARD

- Random drift assemblies are homogenous → they have small Euclidean norm.
  - **X** == [100, 0, 0, 0, 0, 0,…,0] ➔

$$|\boldsymbol{X}| = \sqrt{100^2} = 100$$

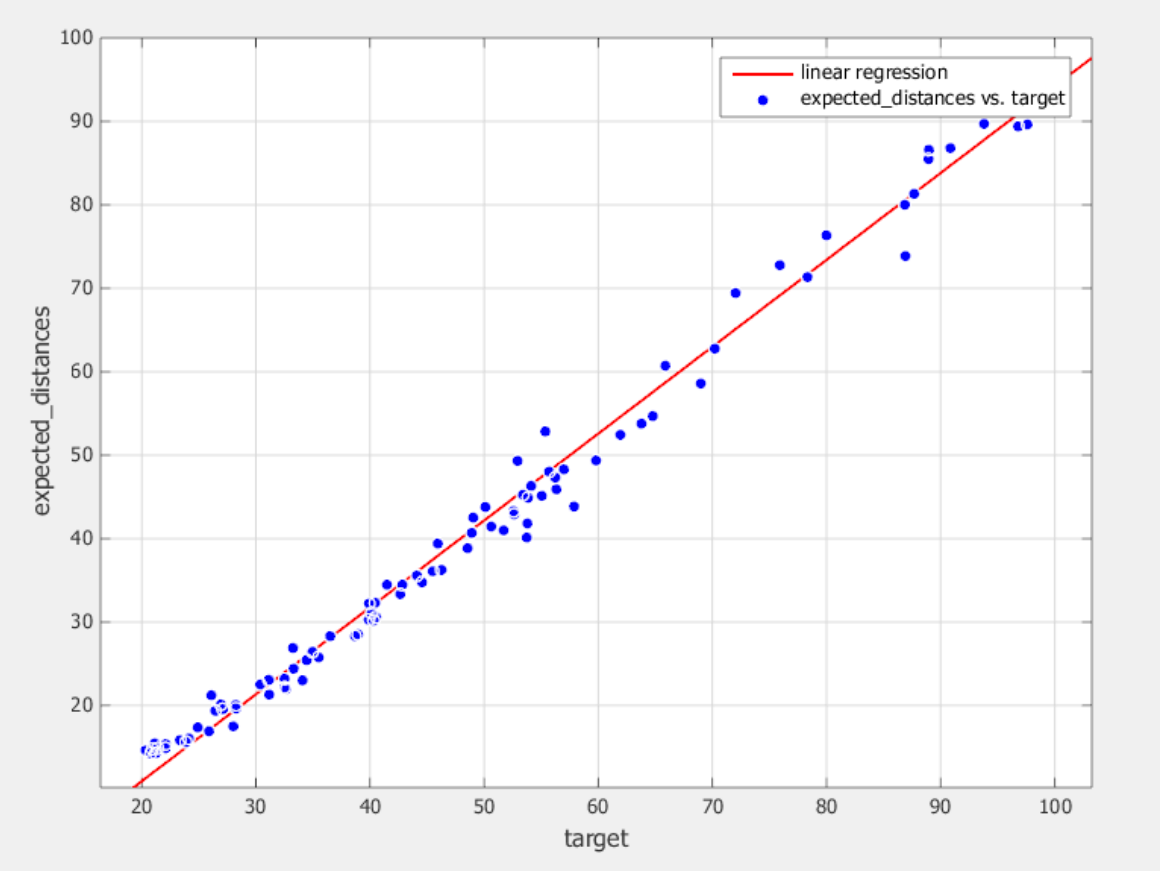
  - **X** == [1, 1, 1, 1, …, 1] ➔

$$|\boldsymbol{X}| = \sqrt{1 + 1 + 1 \ldots} = \sqrt{100} = 10$$

- Distances to compotypes / eigenvectors then depend mostly on the size of the compotype / eigenvector.

Error Catastrophe in GARD

- Indeed, not that bad correlation.
  - $y = 1.041x - 9.89$; $R^2 = 0.9899$



Error Catastrophe in GARD

- K_f and K_b are similar to q
- Of course, there are differences.
  - No complementary replication
  - In this case, what is the master sequence?
  - Back mutation?

Conclusions

- GARD constant population models give distance distributions that are similar to those generated by the quasispecies model.

- GARD replication fidelity shows sensitivity to k_f and k_b. Low k_f results in loss of compotype dominance, just like low q-0.5 results in loss of master sequence.

∴ Compotypes/composomes behave similarly to quasispecies.