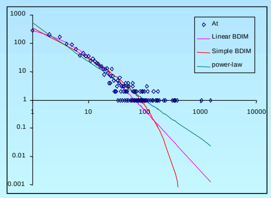
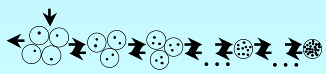
Birth, Death and Innovation Models (BDIM) of genome evolution

Eugene V. Koonin
NCBI, NLM, NIH, Bethesda, MD



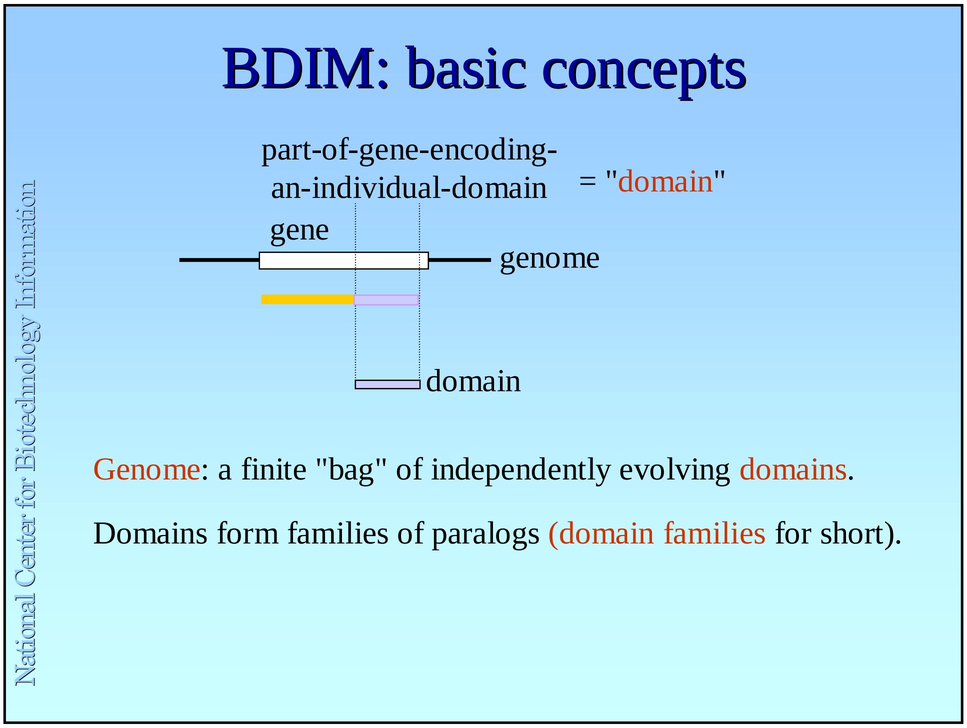


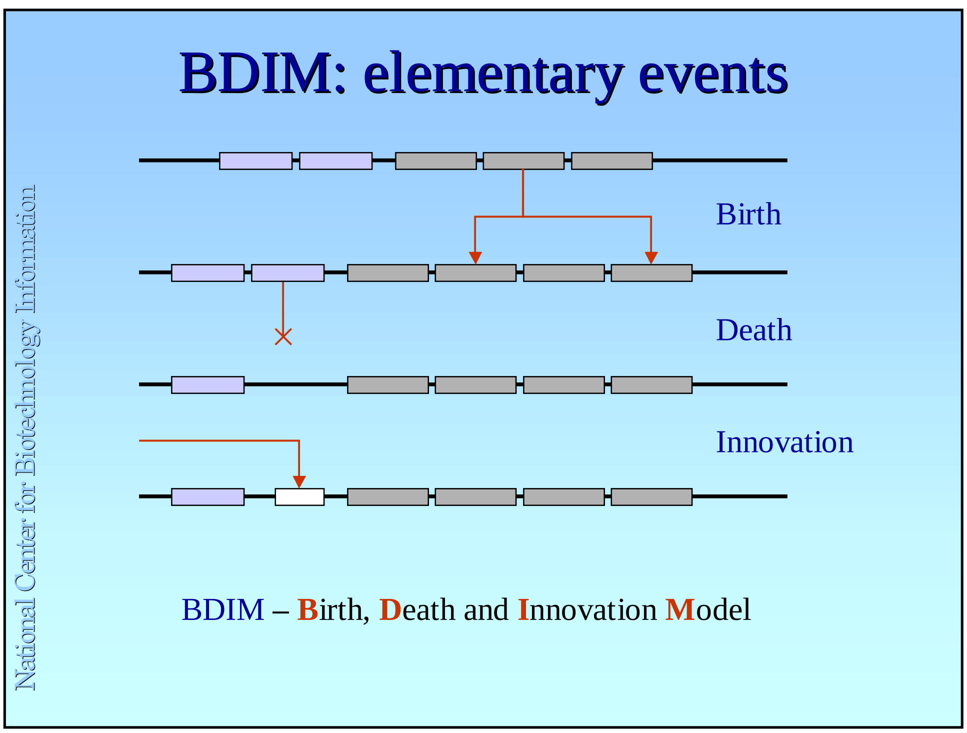
$$df_i(t)/dt = +\lambda_{i-1}f_{i-1}-\delta_i f_i-\lambda_i f_i+\delta_{i+1} f_{i+1}$$

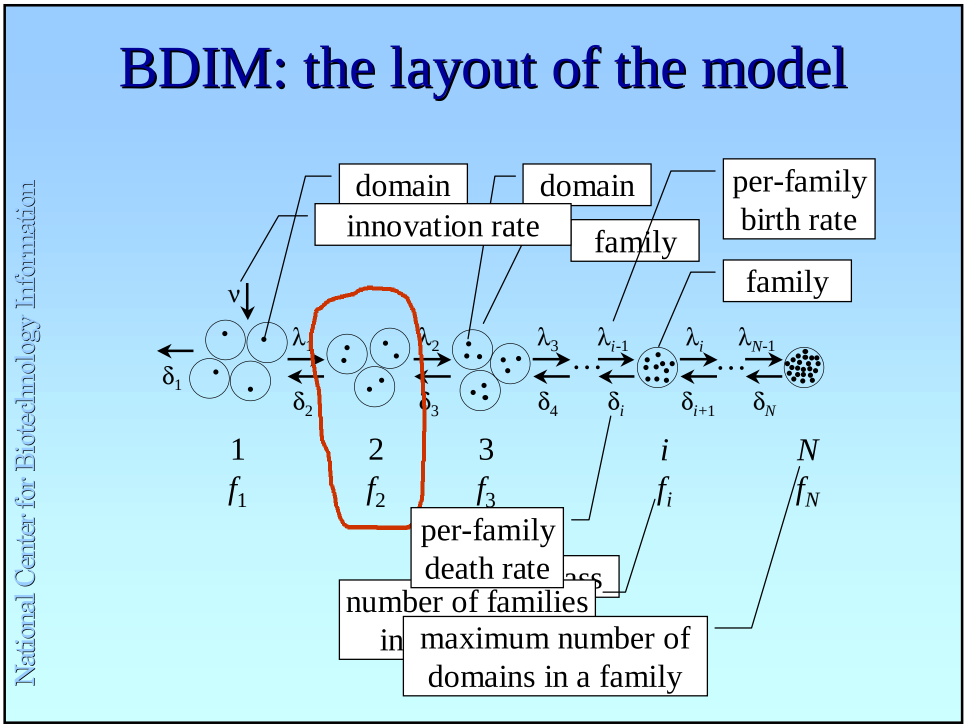
National Center for Biotechnology Information

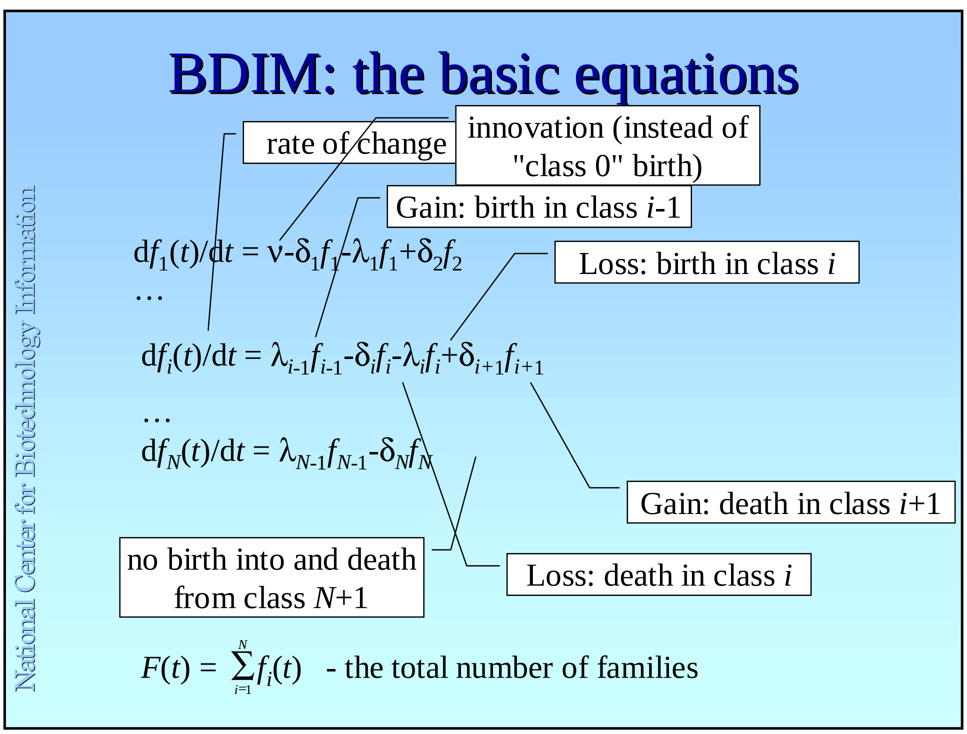
-mathematical formulation and classification of models
-where do the power laws come from
-empirical data fit
-do BDIM tell us anything biologically non-trivial?

National Center for Biotechnology Information


BDIM: basic concepts
part-of-gene-encoding-
an-individual-domain
= "domain"
gene
genome
domain
Genome: a finite "bag" of independently evolving domains.
Domains form families of paralogs (domain families for short).
National Center for Biotechnology Information


BDIM: elementary events
Birth
Death
Innovation
BDIM – Birth, Death and Innovation Model
National Center for Biotechnology Information


BDIM: the layout of the model
domain
innovation rate
domain
family
per-family
birth rate
family
ν
λ1
λ2
λ3
λi-1
λi
λN-1
δ1
δ2
δ3
δ4
δi
δi+1
δN
1
2
3
i
N
f1
f2
f3
fi
fN
per-family
death rate
number of families
maximum number of
domains in a family
National Center for Biotechnology Information


BDIM: the basic equations
rate of change
innovation (instead of
"class 0" birth)
Gain: birth in class i-1
Loss: birth in class i
df1(t)/dt = ν-δ1f1-λ1f1+δ2f2
…
dfi(t)/dt = λi-1fi-1-δifi-λifi+δi+1fi+1
…
dfN(t)/dt = λN-1fN-1-δNfN
Gain: death in class i+1
no birth into and death
from class N+1
Loss: death in class i
F(t) = Σ fi(t) - the total number of families
National Center for Biotechnology Information

BDIM: the equilibrium

$df_i(t)/dt = 0$ - equilibrium for the number of domain families in each size class

$dF(t)/dt = 0$ - equilibrium for the total number of families

•There exists a **unique** and **stable** equilibrium state $f_1, f_2, \ldots, f_N$

•The model reaches equilibrium **exponentially:**

$|f_i(t) - f_i| \sim e^{-kt}$

•The model is "open" at one end only (class 1 families).

A simple condition describes the equilibrium for the **total** number of families:

$$\nu = \delta_1 f_1$$

innovation → ν; Death of **families** → $\delta_1 f_1$

Simple BDIM

Independence hypothesis:

i. all elementary events are independent of each other

ii. the rates of **individual domain** birth (λ) and death (δ) do not depend on *i* (number of domains in a family).

Corollary:

$\lambda_i = \lambda i$

$\delta_i = \delta i$

The basic equation for domain family evolution:

$$df_i(t)/dt = \lambda(i-1)f_{i-1} - (\delta+\lambda)i f_i + \delta(i+1)f_{i+1}$$

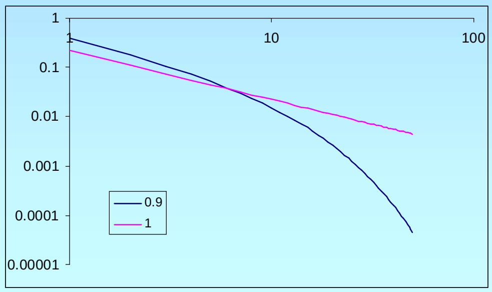
Simple BDIM: equilibrium

The equilibrium solution:

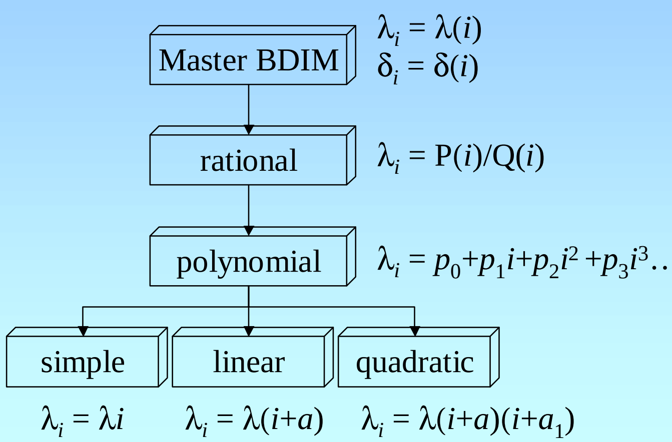
$f_i \sim (\lambda/\delta)^i/i$ - truncated logarithmic distribution

If $\lambda = \delta$

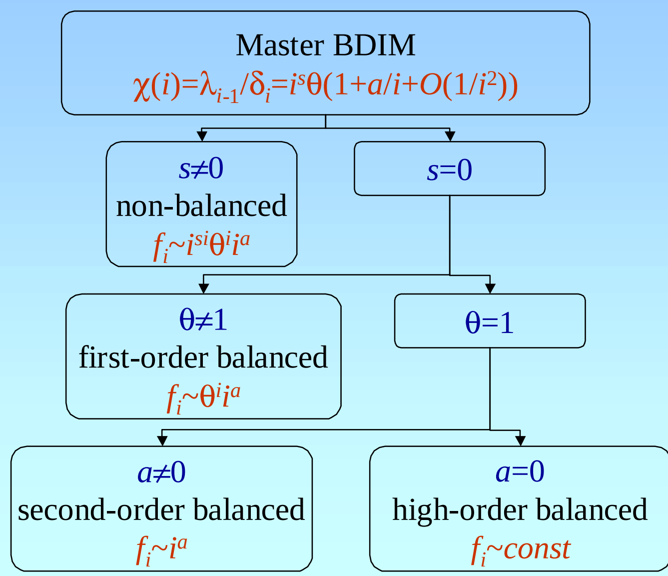
$f_i \sim 1/i$ - **power law (degree = -1)**



BDIM hierarchy



BDIM: balanced and non-balanced models



National Center for Biotechnology Information

BDIM: only first/second order balanced models make sense

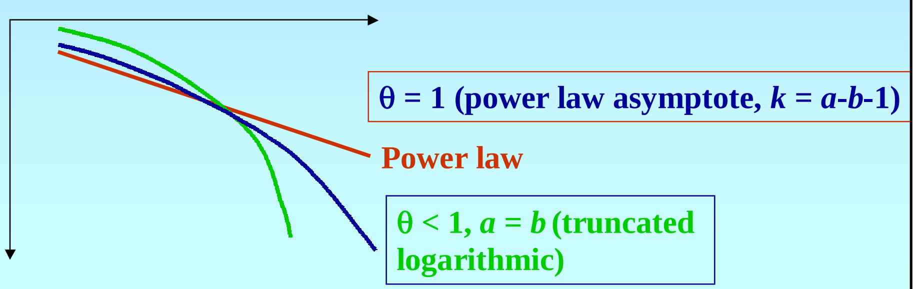
•Non-balanced BDIM: unrealistic family size distributions with extremely strong dependence on i:
either no large families at all or mostly large families
•High-order balanced BDIM: equally unrealistic, uniform distribution

National Center for Biotechnology Information

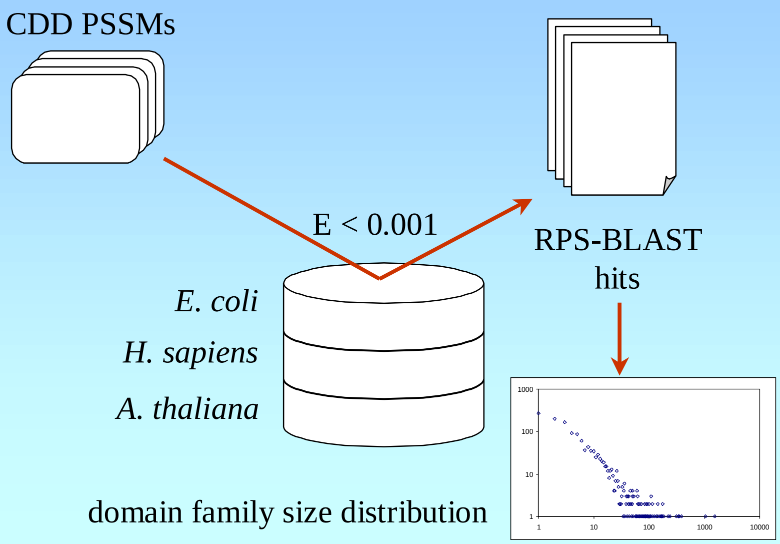
Linear BDIM

$\lambda_i = \lambda(i+a)$
$\delta_i = \delta(i+b)$
$\theta = \lambda/\delta$

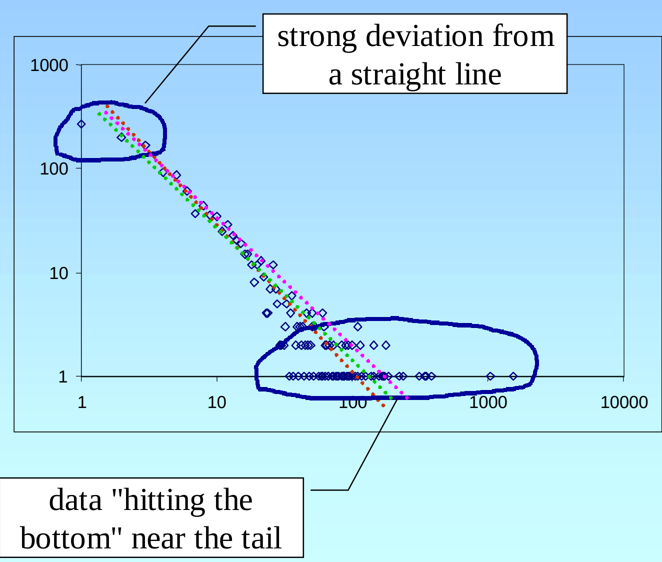
$$f_i = \underbrace{[\Gamma(1+b)/\Gamma(1+a)](\nu/\delta)}_{\text{constant}}\underbrace{\theta^{i-1}}_{\text{exponent}}\underbrace{[\Gamma(i+a)/\Gamma(i+b+1)]}_{\sim i^{a-b-1} \text{ for large } i}$$



Empirical Data: Domains in Genomes

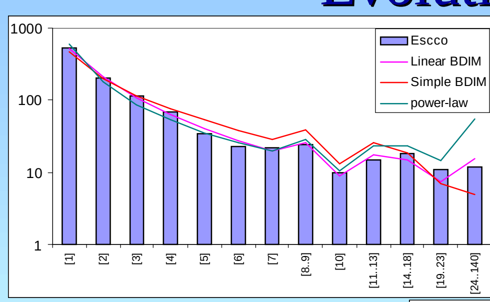


Data Fitting

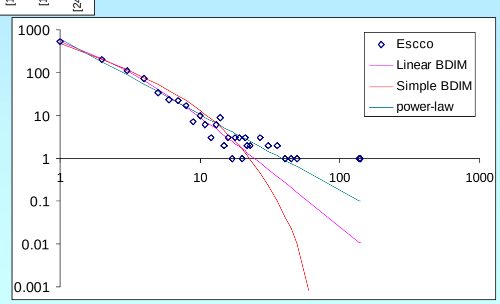


National Center for Biotechnology Information

Data Fitting: Choosing the Model of Evolution

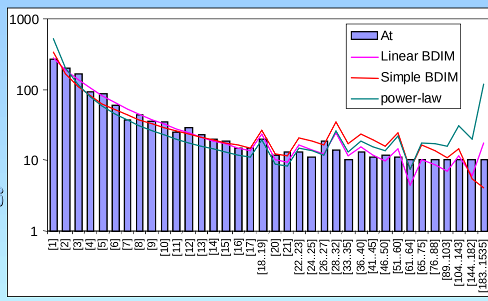


E. coli
Linear BDIM: $P(\chi^2) = 0.71$
Simple BDIM: $P(\chi^2) = 3\times10^{-6}$
Power law: $P(\chi^2) = 2\times10^{-10}$

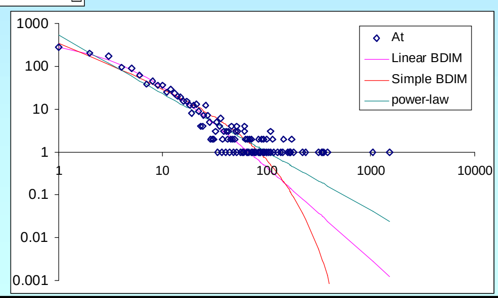


National Center for Biotechnology Information

Data Fitting: Choosing the Model of Evolution



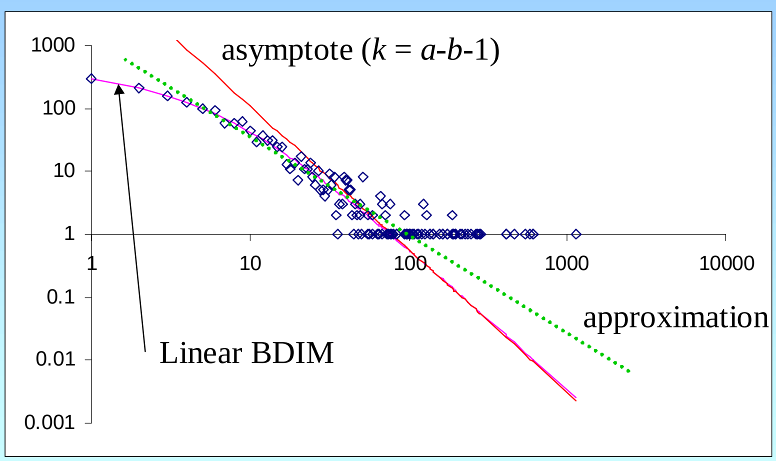
A. thaliana
Linear BDIM: $P(\chi^2) = 0.09$
Simple BDIM: $P(\chi^2) = 6\times10^{-13}$
Power law: $P(\chi^2) = 2\times10^{-54}$



National Center for Biotechnology Information

Data Fitting:

Power Approximation vs Power Asymptote under the linear BDIM



National Center for Biotechnology Information

Linear BDIM parameters and other relevant numbers for bacteria, archaea and eukaryotes

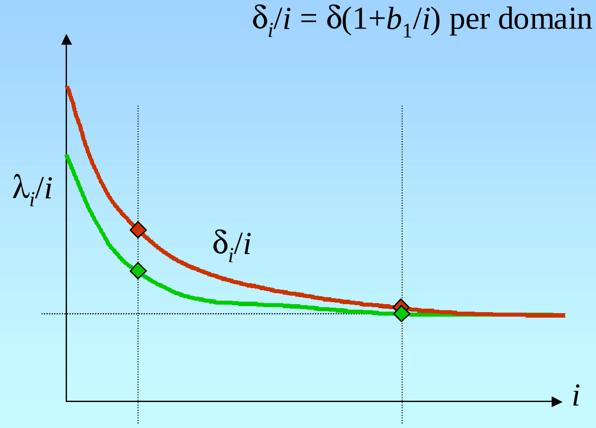
	nORF	nFam	nDom	nHit	*i*Max	d_1 (d'_1)	*a*	*b*	*k*	ν/δ=ν/λ	G
Sc	6340	1080	4575	3331	130	420 (436)	1.55	3.27	-2.72	1861.8	[3.28..3.53]
Dm	13605	1405	11734	7262	335	426 (435)	1.62	2.79	-2.17	1648.2	[8.44..15.50]
Ce	20524	1418	17054	11090	662	423 (421)	1.13	2.03	-1.89	1273.0	[16.18..∞]
At	25854	1405	21238	15006	1535	270 (277)	3.80	4.98	-2.18	1657.7	[17.09..26.89]
Hs	39883	1681	27844	16755	1151	298 (288)	5.16	6.43	-2.27	2136.2	[17.14..22.88]
Thema	1846	772	1683	1268	97	501 (499)	0.14	2.22	-3.08	1606.4	[1.04..1.06]
Metth	1869	693	1480	1150	43	438 (436)	0.12	2.00	-2.88	1305.3	[1.18..1.27]
Sulso	2977	695	1950	1614	81	386 (385)	0.36	2.04	-2.68	1167.8	[1.83..2.00]
Bacsu	4100	1002	3413	2502	124	507 (510)	0.48	2.01	-2.53	1534.6	[2.46..2.79]
Escco	4289	1078	3624	2765	140	523 (519)	0.84	2.54	-2.70	1837.0	[2.45..2.61]

G=total duplication rate/innovation rate

Linear BDIM: Size Does Matter?

$\lambda_i/i = \lambda(1+a_1/i)$ per domain birth rate

$\delta_i/i = \delta(1+b_1/i)$ per domain death rate



National Center for Biotechnology Information

Conclusions

National Center for Biotechnology Information

I. **Only balanced BDIM produce reasonable equilibrium distributions of domain family size; equilibrium is reached rapidly, suggesting a "punctuated equilibrium"-like mode of genome evolution.**

II. **The simplest evolutionary model that adequately describes the observed distribution of domain family size is the linear, second-order balanced BDIM; accordingly, per-domain birth/death rate depends on family size, the larger families being less dynamic in evolution.**

III. **The rates of domain innovation and birth are comparable.**

National Center for Biotechnology Information

The original version of BDIM is fully deterministic.

In order to be able to explore the dynamics of genome evolution, we introduce a stochastic (Markov) version.

National Center for Biotechnology Information

Markov version of BDIM (0 class introduced)

(innovation interpreted as extraction from class 0)

$d\, p_0(t)/dt = -\lambda_0\, p_0(t) + \delta_1\, p_1(t)$,

$d\, p_1(t)/dt = \lambda_0\, p_0(t) - (\lambda_1 + \delta_1)\, p_1(t) + \delta_2 p_2(t)$,

…………………………………………………………………………

$d\, p_i(t)/dt = \lambda_{i-1} p_{i-1}(t) - (\lambda_i + \delta_i)\, p_i(t) + \delta_{i+1} p_{i+1}(t)$ for $1 < i < N$,

…………………………………………………………………………

$d\, p_N(t)/dt = \lambda_{N-1} p_{N-1}(t) - \delta_N\, p_N(t)$

Modified Markov version of BDIM (no 0 class, class 1 immortal)

$d\, p_1(t)/dt = -\lambda_1\, p_1(t) + \delta_2 p_2(t)$,

…………………………………………………………………………

$d\, p_i(t)/dt = \lambda_{i-1} p_{i-1}(t) - (\lambda_i + \delta_i)\, p_i(t) + \delta_{i+1} p_{i+1}(t)$ for $1 < i < N$,

…………………………………………………………………………

$d\, p_N(t)/dt = \lambda_{N-1} p_{N-1}(t) - \delta_N\, p_N(t)$.

no innovation

National Center for Biotechnology Information

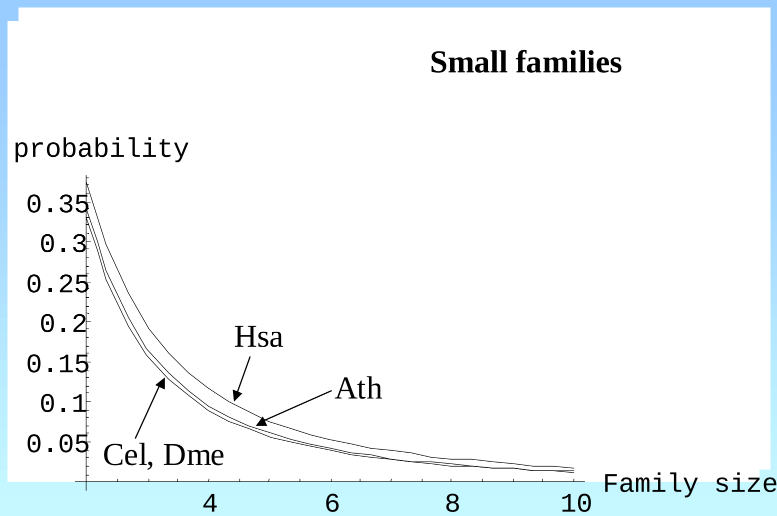
Probability for a family to reach size *n* from size *i* before extinction (size 0)

$$P(i;n) = \left(1 + \sum_{j=1}^{i-1} \prod_{k=1}^{j} \delta_k/\lambda_k\right) / \left(1 + \sum_{j=1}^{n-1} \prod_{k=1}^{j} \delta_k/\lambda_k\right)$$

And, for 2nd order balanced linear BDIM and i=1,

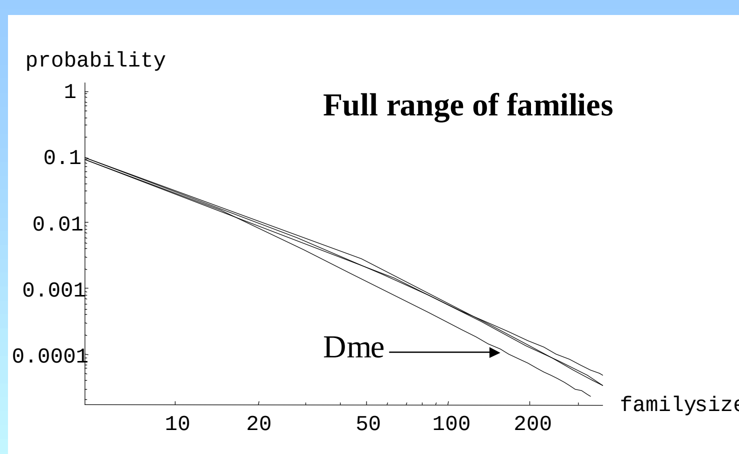
$$P(1,n) = 1/\left(1 + \frac{\Gamma(1+a)}{\gamma\Gamma(1+b)}\left(\frac{\Gamma(b+n+1)}{\Gamma(a+n)} - \frac{\Gamma(2+b)}{\Gamma(1+a)}\right)\right)$$

National Center for Biotechnology Information



Probability that a singleton forms a family with *i* members via a series of duplications before dying (double-log scale) under 2nd order balanced linear BDIM

National Center for Biotechnology Information



Probability that a singleton forms a family with N members via a series of duplications before dying (double-log scale) under 2nd order balanced linear BDIM

National Center for Biotechnology Information

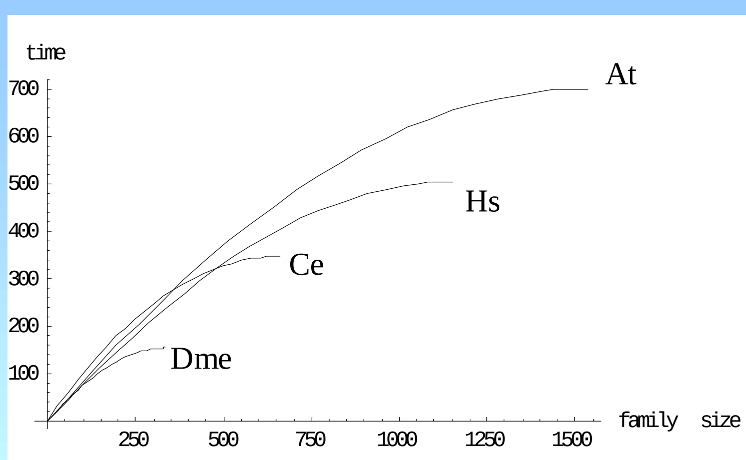
Time before extinction of a family of size *n* for 2nd order balanced linear BDIM

$$E(n) = 1/\lambda \sum_{j=1}^{n} \frac{\Gamma(j+b)}{\Gamma(j+a)} \sum_{k=j}^{N} \frac{\Gamma(k+a)}{\Gamma(k+1+b)}$$

For the time being, we use 1/λ as a natural time scale for BDIM…

National Center for Biotechnology Information

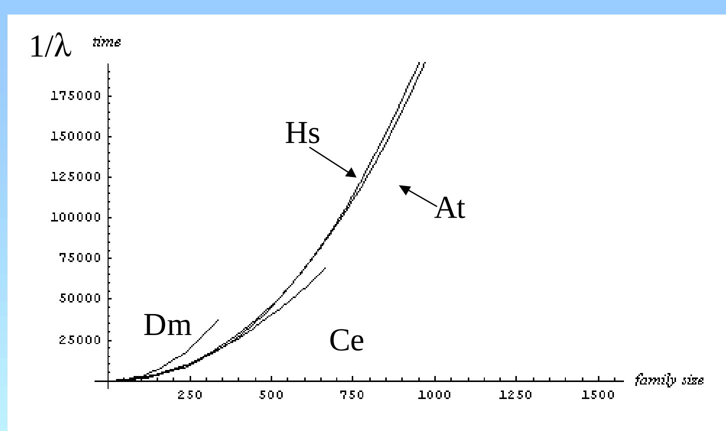
1/λ



Mean time of extinction *E*(*n*) depending on family size for 2nd order balanced linear BDIM

Time required for a singleton to reach family size n under 2nd order balanced linear BDIM (measured in 1/λ units)

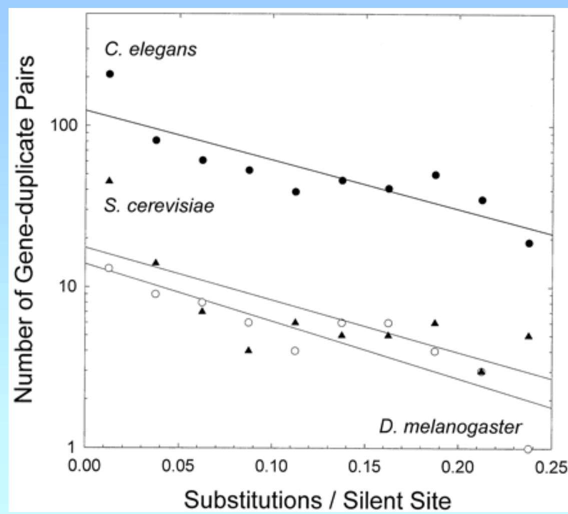
$$M(1;n)=1/\lambda \sum_{k=1}^{n-1} \left(\frac{\Gamma(b+k+1)}{\Gamma(a+k+1)} \sum_{i=1}^{k} \frac{\Gamma(a+i)}{\Gamma(b+i+1)} \right). \qquad (8.7)$$



Mean time for the formation of families depending on *n* under the 2nd order balanced linear BDIM

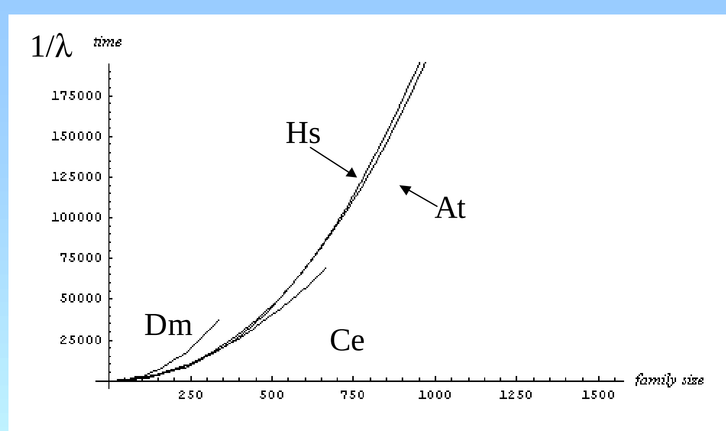
Lynch M, Conery JS. Science 2000 Nov 10;290(5494):1151-5

The Evolutionary Fate and Consequences of Duplicate Genes



0.002 to 0.02 duplications per gene per million years

National Center for Biotechnology Information



Mean time for the formation of families depending on *n* under the 2nd order balanced linear BDIM

Assuming duplication rate 2×10^{-8}/yr,
T(1,N) ~ 10^{13} yr, i.e., ~1,000 x age of Universe

National Center for Biotechnology Information

Stochastic characteristics of the linear BDIM:
i) extremely large difference between times of formation and extinction of the largest families for some genomes – extinction happens **much** faster;
ii) Under the available empirical estimates of duplication rate, the time required for the formation of the largest families is **unrealistically** long

Consequently, we must **replace the linear BDIM** with another model such that:

1) the stationary distribution of the family sizes is the same as for the linear BDIM;
2) the new model provides for much more rapid evolution of gene families under realistic values of duplication and deletion rates;
3) the ratio of family formation and extinction mean times must be significantly less than for the linear BDIM.

National Center for Biotechnology Information

To obtain a new BDIM without changing the stationary distribution:

$\lambda^*_i \rightarrow \lambda_i g(i),\ \ \delta^*_i \rightarrow \delta_i\, g(i\text{-}1)$

$g>0$
$g(0)=1$

Probability of formation of family of size *n* prior to exinction:

$$P^*(1,n) = 1/(1+\frac{\Gamma(1+a)}{\Gamma(1+b)}\sum_{k=1}^{n-1}\frac{1}{g(k)}\frac{\Gamma(b+k+1)}{\Gamma(a+k+1)});$$

Extinction time:

$$E^*_s = 1/\lambda\sum_{k=1}^{s}\ \sum_{i=k}^{N} 1/g(k\text{-}1)\ [\frac{\Gamma(a+i)}{\Gamma(b+i+1)}\ \frac{\Gamma(b+k)}{\Gamma(a+k)}]$$

Family formation time (class 1 immortal):

$$M^*(1;n) = 1/\lambda\sum_{k=1}^{n-1}\frac{1}{g(k)}\frac{\Gamma(b+k+1)}{\Gamma(a+k+1)}\sum_{i=1}^{k}\frac{\Gamma(a+i)}{\Gamma(b+i+1)}.$$

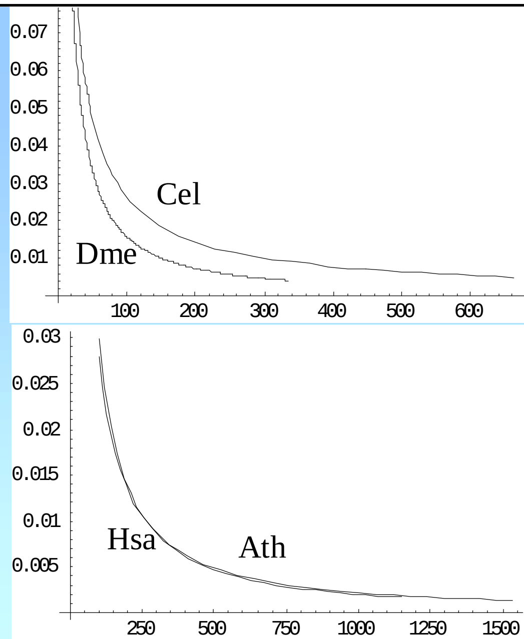
National Center for Biotechnology Information

Quadratic BDIM

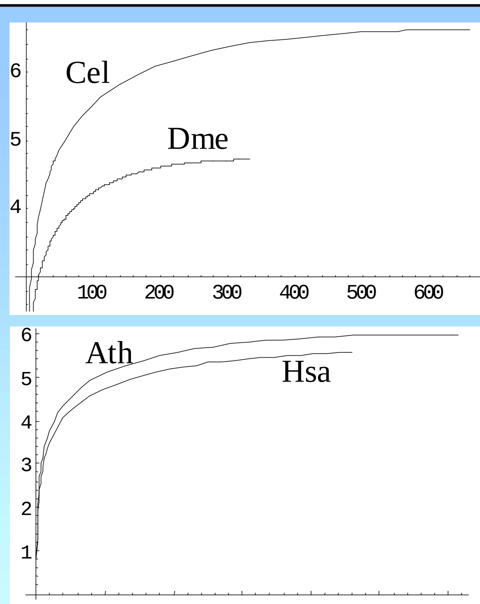
Let us consider the 2nd order balanced polynomial (quadratic) Markov BDIM with duplication and deletion rates

$\lambda_i = \lambda(i+a)(i+1),$
$\delta_i = \lambda(i+b)i.$

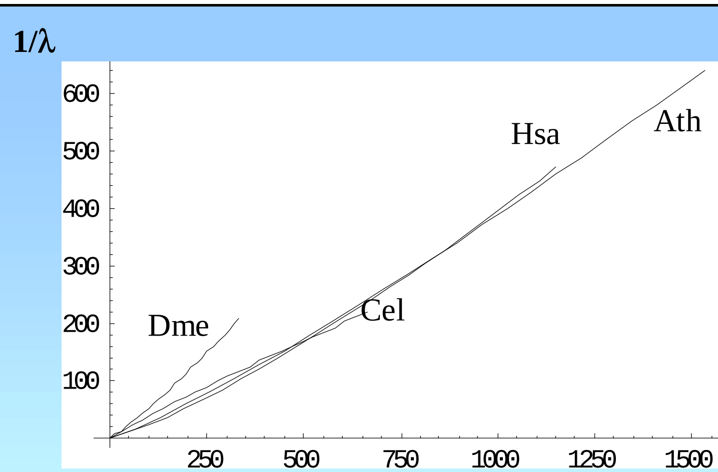
This may be interpreted as introducing pairwise interactions between family members



Probability that a singleton forms a family with *i* members via a series of duplications before dying under the quadratic BDIM

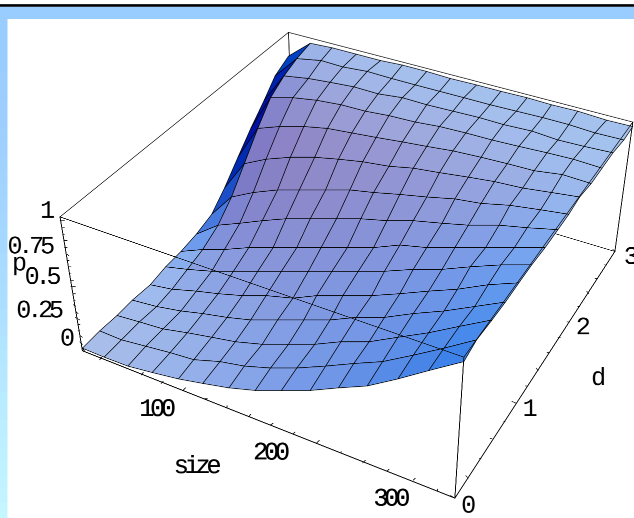


Mean time of extinction under the quadratic BDIM, $E_2(n)$, depending on family size *i*

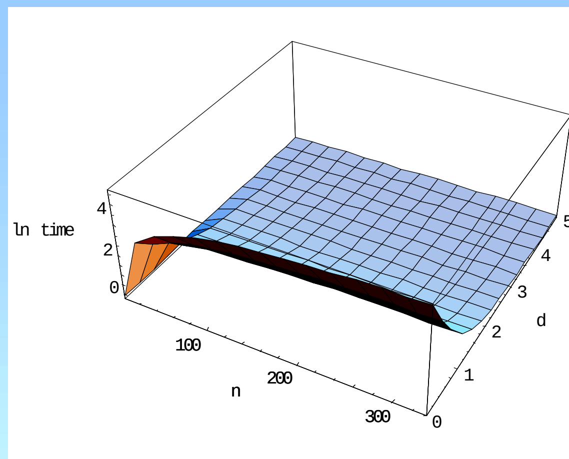


Mean time of formation of a family with *i* members under quadratic BDIM

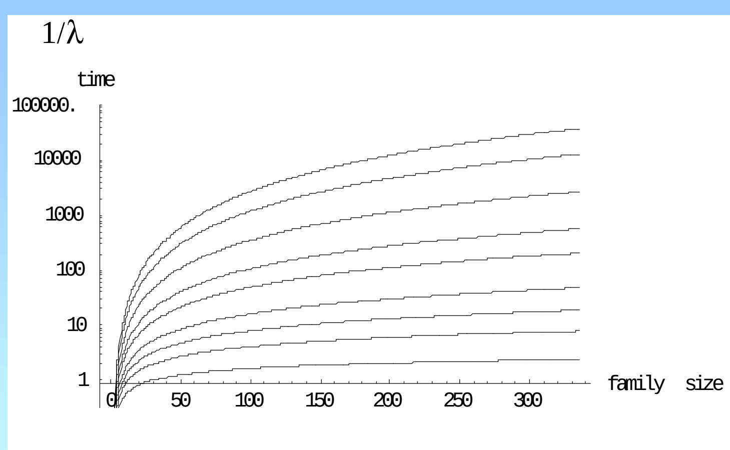
Assuming duplication rate (λ) ~$2x10^{-8}$/yr, T(1,N) ~ 10^{10} yr – about the age of Universe, still longer than real evolution times but not by much



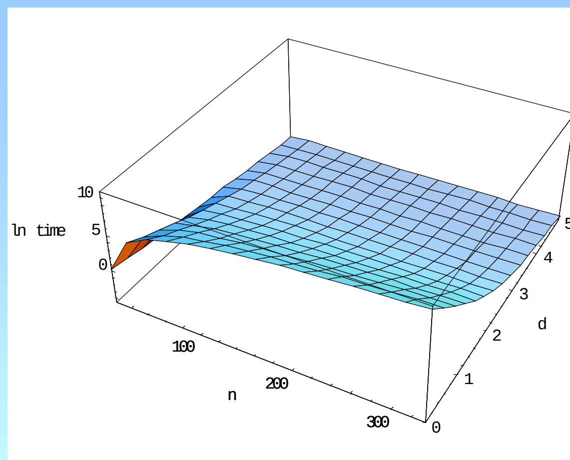
Probability $P_{d+1}(i,N)$ of formation of the largest family before extinction depending on the model degree $k=1+d$ and the initial family size; the values of parameters $a=1.55$, $b=2.79$, $N=335$ are for *Dme*.



The dependence of $ln(E_{d+1}(n))$ – extinction time - on family size, $1<n<335$, and on the model degree, $0<d<5$;
the values of parameters a=1.64, b=2.79, N=335 are for *Dme.*

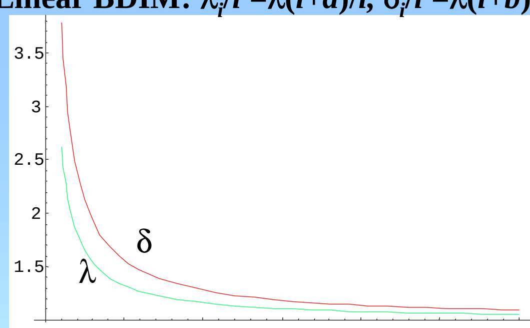


Mean time of family formation, $M_{d+1}(n)$ (log scale) for BDIM of different degrees (from linear to cubic):
d=0, 0.2, 0.5, 0.8, 1., 1.3, 1.5, 1.7, 2 (from top to bottom);
the parameter values are for *Dme*

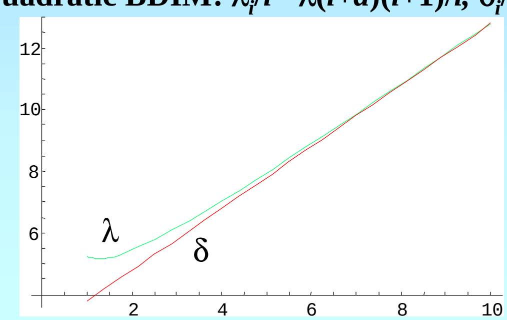


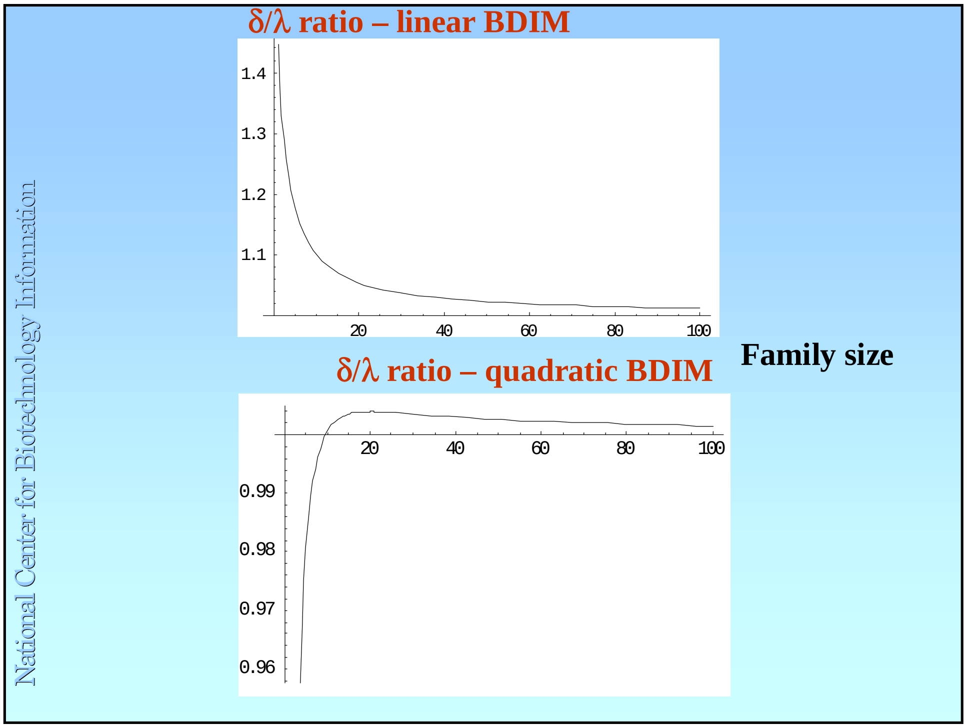
Mean time of family formation $M_{d+1}(n)$ depending on initial of family size ($1<n<335$) and model degree d; the parameter values a=1.62, b=2.79, N=335 are for *Dme*

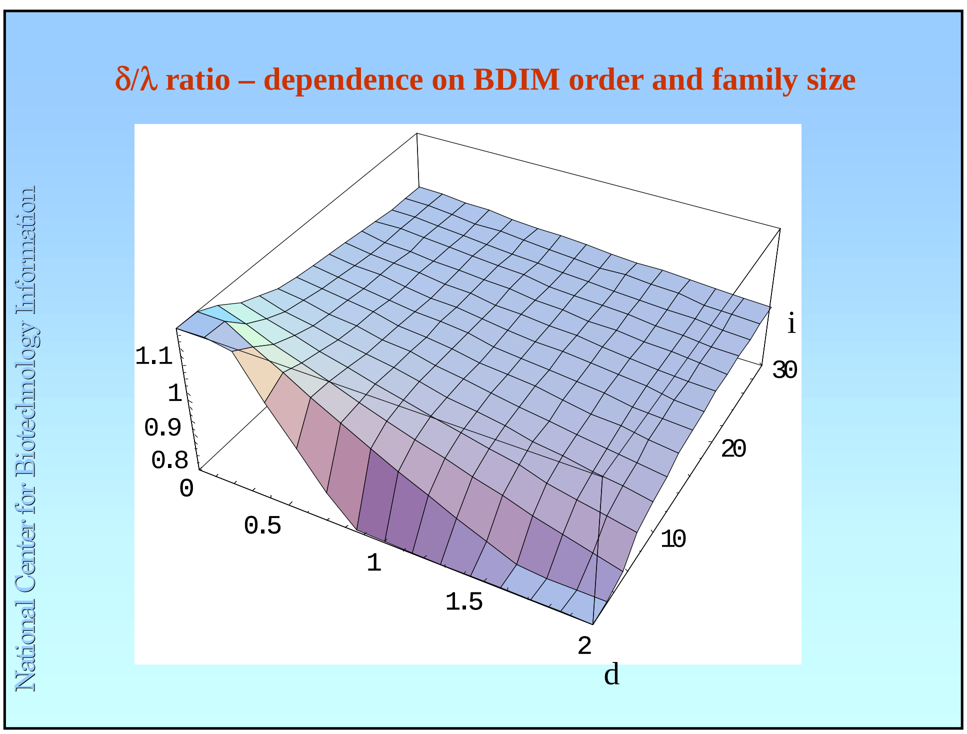
Linear BDIM: $\lambda_i/i = \lambda(i+a)/i$, $\delta_i/i = \lambda(i+b)/i$.



Quadratic BDIM: $\lambda_i/i = \lambda(i+a)(i+1)/i$, $\delta_i/i = \lambda(i+b)$.




δ/λ ratio – linear BDIM
1.4
1.3
1.2
1.1
20
40
60
80
100
Family size
δ/λ ratio – quadratic BDIM
20
40
60
80
100
0.99
0.98
0.97
0.96
National Center for Biotechnology Information


δ/λ ratio – dependence on BDIM order and family size
1.1
1
0.9
0.8
0
0.5
1
1.5
2
d
i
30
20
10
National Center for Biotechnology Information

Conclusions on stochastic (Markov) BDIMs and some general conclusions

National Center for Biotechnology Information

2nd order balanced linear BDIM is sufficient to explain the observed distributions of domain family size

However, when a stochastic model is used to estimate the time required to reach the maximum family size, the estimate is ~10,000 times greater than the time suggested by empirical data

BDIMs of the degree 2-3 (quadratic/cubic) formally solve the problem by predicting evolutionary rate compatible with observations

Conclusions on stochastic (Markov) BDIMs and some general conclusions

National Center for Biotechnology Information

Biological interpretation of “interaction” between paralogs, which is intrinsic in higher-order BDIMs – does it reflect selection?

The dependence of birth and death rates on family size dramatically changes depending on BDIM order. This needs to be tested against detailed empirical analysis of paralogous families.

National Center for Biotechnology Information

Some future directions…

- **Incorporating selection into BDIM**
- **Combining BDIM with phylogenetic tree analysis**
- **Modeling evolutionary processes that lead to similar distributions in other contexts, e.g., multidomain protein architectures and interaction networks**

National Center for Biotechnology Information

Tomorrow??

Genomics today

"***There are two kinds of science: physics and stamps collection***"

Attributed to Ernest Rutherford

Acknowledgments

Georgy Karev (NCBI, NIH),
Yuri Wolf (NCBI, NIH),
Andrey Rzhetsky (Columbia University),
Faina Berezovskaya (Howard University)

National Center for Biotechnology Information