

DETEKCE

SELEKCE

Testování neutrality vs. selekce

základní mírou relativního významu selekce a driftu je poměr dN/dS (ω)

dN (d_N, K_a) = průměrný počet nukleotidových rozdílů mezi sekvencemi
na 1 nesynonymní pozici

měří míru rozdílnosti 2 homologních kódujících sekvencí z hlediska aminokyselin, tj. do jaké míry se liší v nesynonymních pozicích

dS (d_S, K_s) = průměrný počet nukleotidových rozdílů mezi sekvencemi
na 1 synonymní pozici

měří míru rozdílnosti 2 homologních kódujících sekvencí z hlediska tichých substitucí, tj. do jaké míry se liší v synonymních pozicích

Výpočet dN/dS :

ACTCCGAACGGGGCGTTAGAGTTGAAACCCGTTA GA
* * * * * * * * **
ACGCCGATCGGCGCGATAGGGTTCAAGCTCGTACGA

substituce

přepsáno do sekvencí aminokyselin:

TPNGALELKPVR

* * * *

nesynonymní
záměny

TPIGAIGFKLVR

tj. 5 nesynonymních substitucí

protože celkový počet záměn je 10 (viz 10 hvězdiček mezi sekvencemi DNA), 5 musí být synonymních

```
ACTCCGAACGGGGCGTTAGAGTTGAAACCCGTTAGA
 *      *      *      *      *      *      *      *
ACGCCGATCGGCGCGATAGGGTTCAAGCTCGTACGA
```



bud' ACT (sekvence 1), nebo ACG (sekvence 2) \Rightarrow změna z A na
kteroukoli bázi způsobí záměnu aminokyseliny (např. CCT, GCT, TCT)
 \Rightarrow pozice 1 je nesynonymní

```
ACTCCGAACGGGGCGTTAGAGTTGAAACCCGTTAGA
 *      *      *      *      *      *      *      *
ACGCCGATCGGCGCGATAGGGTTCAAGCTCGTACGA
```

Pozice 2: podle genetického kódu každá substituce na 2. místě kodonu
je nesynonymní

ACTCCGAACGGGGCGTTAGAGTTGAAACCCGTTAGA
* * * * * * * * * *
ACGCCGATCGGCGCGATAGGGTTCAAGCTCGTACGA

Pozice 3: 4 potenciální aminokyseliny lišící se ve 3. pozici – ACT, ACG, ACC, ACA → všechny kódují stejnou aminokyselinu (threonin, T) ⇒ všechny substituce jsou synonymní ⇒ pozice 3 je synonymní (tato pozice je 4-násobně degenerovaná)

pozice 4 (C v CCG): všechny substituce nesynonymní

pozice 5: nesynonymní atd.

ACTCCGAACGGGGCGTTAGAGTTGAAACCCGTTAGA
* * * * * * * * * *
ACGCCGATCGGCGCGATAGGGTTCAAGCTCGTACGA

Pozice 9:

ACTCCG AAC GGGCGTTAGAGTTGAAACCCGTTAGA
 * * * * * * * * * *
 ACGCCGATC GGC GCGATAGGGTTCAAGCTCGTACGA

Pozice 9: v sekvenci 1 = 3. pozice kodonu AAC (asparagin, N),
 v sekvenci 2 kodon ATC (isoleucin, I)

mutace v AAC → AAT (asparagin, N), AAG, AAA (obě lysin, K) ⇒
 C = 2-násobně degenerovaná → 9. pozice z 1/3 synonymní a
 ze 2/3 nesynonymní

podobně mutace C v ATC → ATT, ATA (obě isoleucin, I), ATG (methionin,
 M) ⇒ 3-násobně degenerovaná pozice – 2/3 synonymní, 1/3
 nesynonymní

⇒ průměr $\frac{1}{2}(\frac{1}{3} \text{ synonymních} + \frac{2}{3} \text{ nesynonymních}) +$
 $\frac{1}{2}(\frac{2}{3} \text{ synonymních} + \frac{1}{2} \text{ nesynonymních})$
 $= \frac{1}{2} \text{ synonymních a } \frac{1}{2} \text{ nesynonymních}$

⇒ pozice 9 je částečně synonymní a částečně nesynonymní

site # 123456789111111111122222222223333333
 012345678901234567890123456

syn 00100100 $\frac{1}{2}$ 001001 $\frac{1}{4}$ 0 $\frac{1}{2}$ 00 $\frac{1}{3}$ $\frac{1}{3}$ 0 $\frac{1}{3}$ 00 $\frac{1}{3}$ 001001 $\frac{1}{3}$ 0 $\frac{2}{3}$ sum = 7.5833

non 11011011 $\frac{1}{2}$ 110110 $\frac{3}{4}$ 1 $\frac{1}{2}$ 11 $\frac{2}{3}$ $\frac{2}{3}$ 1 $\frac{2}{3}$ 11 $\frac{2}{3}$ 110100 $\frac{2}{3}$ 1 $\frac{1}{3}$ sum = 28.4167

$$dN = \frac{\text{poč. nsyn substitucí}}{\text{poč. nsyn pozic}} = \frac{5}{28,417} = 0,176$$

$$dS = \frac{\text{poč. syn substitucí}}{\text{poč. syn pozic}} = \frac{5}{7,583} = 0,659$$

$$\frac{dN}{dS} = \frac{0,176}{0,659} = 0,269$$

Interpretace dN/dS :

1. všechny nesynonymní substituce jsou neutrální:

počet synonymních i nesynonymních neutrálních mutací
fixovaných každou generací = μ
 $\Rightarrow dN/dS = \mu/\mu = 1$

2. část nesynonymních substitucí je neutrálních, zbytek škodlivých:

$$dS = \mu$$

v každé generaci fixace f neutrálních nesynonymních mutací
 $\Rightarrow (1 - f)$ škodlivých mutací se nezafixuje

$$dN = f\mu + (1 - f)0 = f\mu$$

$$dN/dS = f\mu/\mu = f$$

Protože f je vždy < 1 , platí $dN/dS < 1$

Závěr: $dN/dS < 1$ indikuje působení purifikující selekce

3. část f mutací je neškodných a $(1 - f)$ škodlivých; z neškodných mutací je část θ prospěšných a $(1 - \theta)$ neutrálních:

$$dS = \mu$$

$(1 - f)$ se nefixuje

$f(1 - \theta)$ neutrálních \Rightarrow fixace frekvencí μ za generaci

$f\theta$ prospěšných, vznik rychlostí $2N\mu$ za generaci, pravděpodobnost fixace rovna selekčnímu koeficientu s

\Rightarrow počet nesynonymních substitucí fixovaných každou generaci:

$$dN = (1 - f)0 + f(1 - \theta)\mu + f\theta 2N\mu s$$

$$\Rightarrow dN/dS = [(1 - f)0 + f(1 - \theta)\mu + f\theta 2N\mu s]/\mu = f(1 - \theta) + f\theta 2Ns$$

$dN/dS > 1$ pokud θ velká, konkrétně

$$\theta > \frac{1 - f}{f} \frac{1}{(2N - s)}$$

Závěr: $dN/dS > 1$ indikuje působení pozitivní selekce

Pozn.: $dN/dS < 1$ nemusí znamenat, že pozitivní selekce nepůsobí, pouze že ji tímto způsobem nemůžeme detekovat

Shrnutí:

1. $dN/dS = 1$: substituce aminokyselin převážně **neutrální** (ale: pozitivní selekce může vyrušit působení selekce purifikující)

Shrnutí:

2. $dN/dS < 1$: **purifikující selekce**

(ale: některé AA mohly být fixovány pozitivní selekcí, purifikující selekce ale silnější)

3. $dN/dS > 1$: **pozitivní selekce** fixovala některé AA, některé substituce mohly být způsobeny driftem (ale: purifikující selekce mohla působit, ale nebyla dost silná, aby převážila nad selekcí pozitivní)

Kromě výpočtu synonymních a nesynonymních pozic a synonymních a nesynonymních substitucí nutná ještě korekce pro opakované substituce na téže pozici

→ pro výpočty nutné zjednodušující předpoklady, navíc nemůžeme přesně zjistit počet opakovaných substitucí

Odhad pomocí maximální věrohodnosti (*maximum likelihood*):
simultánní odhad všech 3 kroků současně

poskytuje navíc odhad doby divergence a poměr T_s/T_v

Ke kvantifikaci počtu substitucí lze:

rekonstruovat ancestrální sekvenci a spočítat změny na jednotlivých pozicích (výsledek bude pravděpodobně podhodnocený)

bayesovský přístup: použít substituční rychlosti (v apriorních kategoriích)

generovat substituční rychlosti pro jednotlivé kodony

Tajimův test neutrality

měření rovnováhy mutace a driftu pomocí heterozygotnosti $\theta = 4N_e\mu$

θ lze odhadovat i jinými způsoby:

π_{ij} = počet párových rozdílů (SNP) mezi sekvencemi i a j
(... celkem $n(n - 1)/2$ možných párových srovnání)

$$\hat{\pi} = \frac{\sum_{i=1}^j \sum_{j=2}^n \pi_{ij}}{\frac{n(n-1)}{2}}$$

suma párových rozdílů

počet párových srovnání

v případě DNA sekvencí dělíme ještě jejich délkou

S = počet segregujících pozic:

$$\Theta = \frac{S}{\sum_{i=1}^{n-1} \left(\frac{1}{i}\right)}$$

$1/1 + 1/2 + \dots + 1/(n-1)$

při modelu nekonečných pozic a neutrální evoluci platí:

$$\hat{\pi} = \Theta$$

Fumio Tajima (1989):
$$D = \frac{\hat{\pi} - \Theta}{\sqrt{\text{Var}(\hat{\pi} - \Theta)}}$$

Př.:

	*	*	*	*
1	ACCCG	AATTC	CAATC	CGGTT
2	AACTG	AATTC	GAATC	CGGTT
3	AACTG	AATTC	CAATC	CGGTT
4	ACCTG	AATTC	TAATC	CGGAT

párová srovnání:

1-2: 3 rozdíly

1-3: 2 rozdíly

1-4: 3 rozdíly

2-3: 1 rozdíl

2-4: 3 rozdíly

3-4: 3 rozdíly

prům. $\pi = (3+2+3+1+3+3)/6 = 2,5$

S = 4 segregující pozice

$$\Theta = 4 / (1/1 + 1/2 + 1/3) = 4 / 1,83 = 2,186 \quad \hat{\pi} - \Theta = 2,5 - 2,186 = 0,314$$

$D < 0$:

nadbytek polymorfismů s nízkou frekvencí vzhledem k teoretickému předpokladu \Rightarrow purifikující selekce, *selective sweep*
(+ populační expanze!)

$D > 0$:

nadbytek polymorfismů s nízkou i vysokou frekvencí vzhledem k předpokladu \Rightarrow balancující selekce (+ redukce populační velikosti!)

Signifikance?

nelze použít klasické P

Tajima (1989): parametrická aproximace beta rozdělením

Hudson (1990): generování náhodných vzorků za předpokladu neutrality a populační stability \rightarrow hodnota $P =$ podíl náhodných výsledků \leq vypočtené D

Value of Tajima's D	Mathematical reason	Biological interpretation 1	Biological interpretation 2
Tajima's $D=0$	Pi equivalent to Theta (Observed=Expected). Average Heterozygosity=# of Segregating sites.	Observed variation similar to expected variation	Population evolving as per mutation-drift equilibrium. No evidence of selection
Tajima's $D<0$	Pi less than Theta (Observed<Expected). Fewer haplotypes (lower average heterozygosity) than # of segregating sites.	Rare alleles present at low frequencies	Recent selective sweep, population expansion after a recent bottleneck, linkage to a swept gene
Tajima's $D>0$	Pi greater than Theta (Observed>Expected). More haplotypes (more average heterozygosity) than # of segregating sites.	Multiple alleles present, some at low, others at high frequencies	Balancing selection, sudden population contraction

McDonaldův-Kreitmanův test

John H. McDonald and Martin Kreitman (1991):

srovnání vnitrodruhového polymorfismu a mezidruhové divergence

D_s = počet synonymních substitucí^{*)} na sekvenci

D_n = počet nesynonymních substitucí na sekvenci

P_s = počet synonymních polymorfních pozic na sekvenci

P_n = počet nesynonymních polymorfních pozic na sekvenci

H_0 : $D_n/D_s = P_n/P_s \Rightarrow$ neutrální evoluce

H_1 : $D_n/D_s \neq P_n/P_s \Rightarrow$ selekce

^{*)} substituce = u 2 druhů fixována odlišná báze

negativní (purifikující) selekce:

škodlivé mutace silně ovlivňují polymorfismus

$D_n/D_s < P_n/P_s$, tj. poměr nesynonymní/synonymní variability mezi druhy je nižší než poměr nesynonymní/synonymní variability uvnitř druhů

pozitivní selekce:

prospěšné mutace se rychle šíří \Rightarrow neovlivňují polymorfismus, ale mají vliv na mezidruhovou divergenci

$D_n/D_s > P_n/P_s$, tj. poměr nesynonymní/synonymní variability mezi druhy je vyšší než poměr nesynonymní/synonymní variability uvnitř druhů

podíl substitucí fixovaných selekcí: $\alpha = 1 - \frac{D_s P_n}{D_n P_s}$

Problémy MKT:

podhodnocení α v důsledku existence mírně škodlivých mutací,
odlišných mutačních rychlostí v různých částech genomu,
proměnlivosti v koalescenčních historiích různých částí genomu,
změn v efektivní velikosti populace

× tyto problémy ale neznamenaají, že MKT považován za nespolehlivý

další potenciální problém: *infinite-sites model*

→ často odchylky od modelu uvnitř druhů, tím větší v mezidruhových srovnáních

Detekce selekce na úrovni kodonů

Které kodony pod pozitivní/negativní selekcí?

substituční model, fylogenetický strom, výpočet dN/dS pro každý kodon
v případě sekvencí složených z více jedinců (např. viry) odhad pozitivní
selekce na úrovni populace

Kdy v minulosti selekce působila?

dN/dS mapováno na jednotlivé větve fylogenetického stromu

Působí selekce uvnitř rekombinujících fragmentů?

např. program Datamonkey (<http://www.datamonkey.org>)

Selekce v heterogenním prostředí

variabilita prostředí:

v prostoru

v čase

v hrubém měřítku: během života jedno prostředí

v jemném měřítku: během života více prostředí

selekce: měkká
tvrdá

1. Prostředí proměnlivé v hrubém měřítku (*coarse-grained environment*) – prostorová variabilita

Leveneho model:

c_i = část celkové populace obývající habitat i

měkká selekce

Př.: svijonožec *Balanus balanoides*

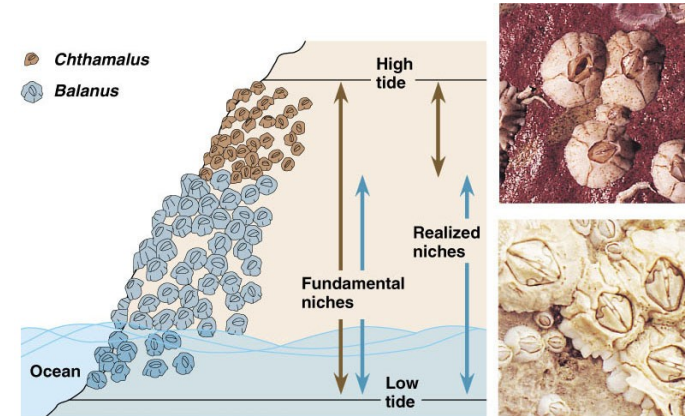


Table 14.1. Habitat-Specific Viability Estimates for *Mpi* Genotypes in Northern Acorn Barnacle

Habitat	<i>SS</i>	<i>SF</i>	<i>FF</i>
Exposed substrate in high intertidal zone	0.696	1	1.424
Exposed substrate in low intertidal zone	0.898	1	1.012
Under algal canopy in high intertidal zone	1.519	1	0.880
Under algal canopy in low intertidal zone	0.913	1	0.976

Source: From Schmidt and Rand (2001).

Note: All values are measured relative to the fitness of the heterozygote.

Table 14.1. Habitat-Specific Viability Estimates for *Mpi* Genotypes in Northern Acorn Barnacle

Habitat	<i>SS</i>	<i>SF</i>	<i>FF</i>
Exposed substrate in high intertidal zone	0.696	1	1.424
Exposed substrate in low intertidal zone	0.898	1	1.012
Under algal canopy in high intertidal zone	1.519	1	0.880
Under algal canopy in low intertidal zone	0.913	1	0.976

$c_1 = 0,4$
 $c_2 = 0,05$
 $c_3 = 0,5$
 $c_4 = 0,05$



Source: From Schmidt and Rand (2001).

Note: All values are measured relative to the fitness of the heterozygote.

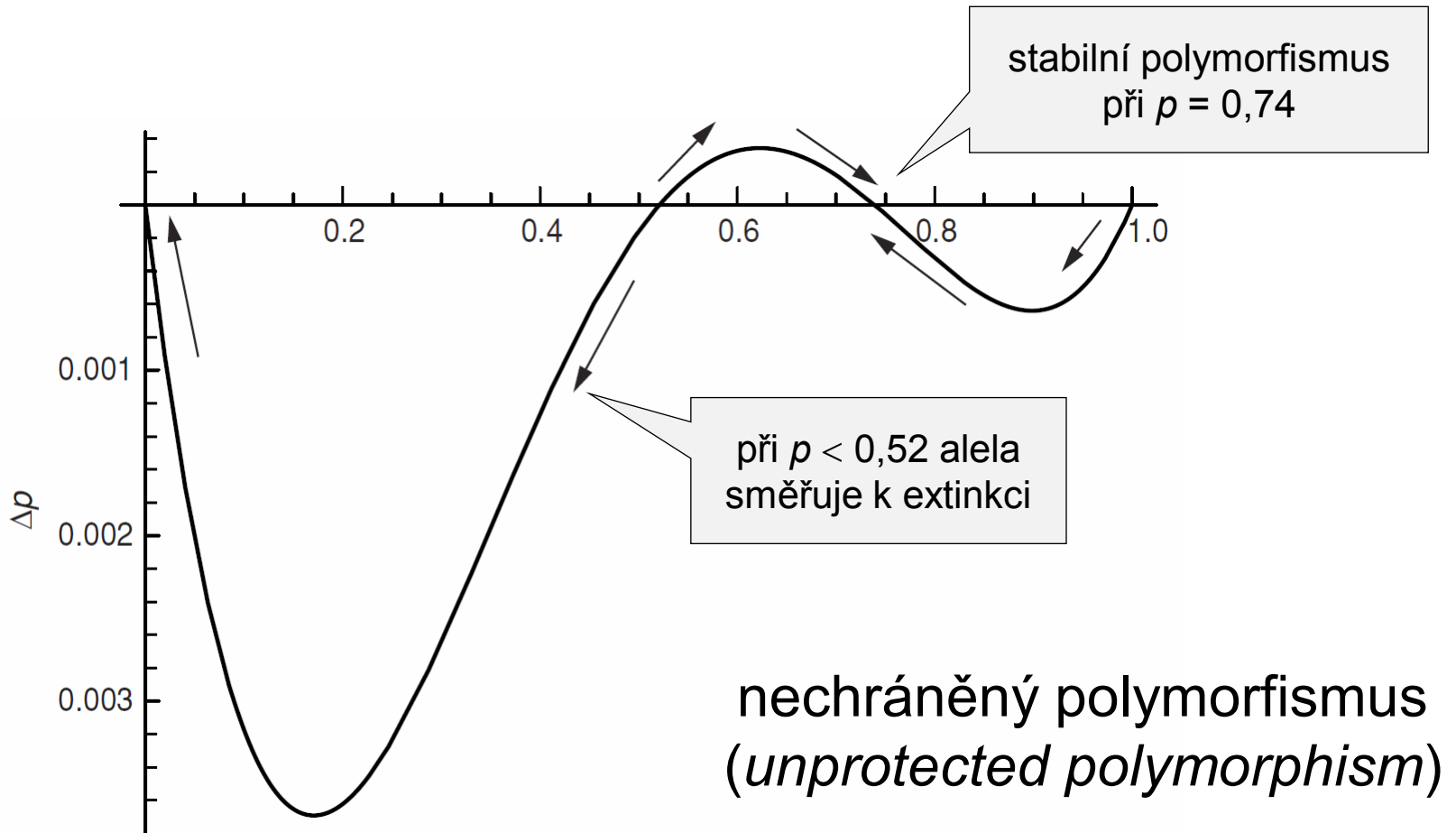
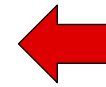


Table 14.1. Habitat-Specific Viability Estimates for *Mpi* Genotypes in Northern Acorn Barnacle

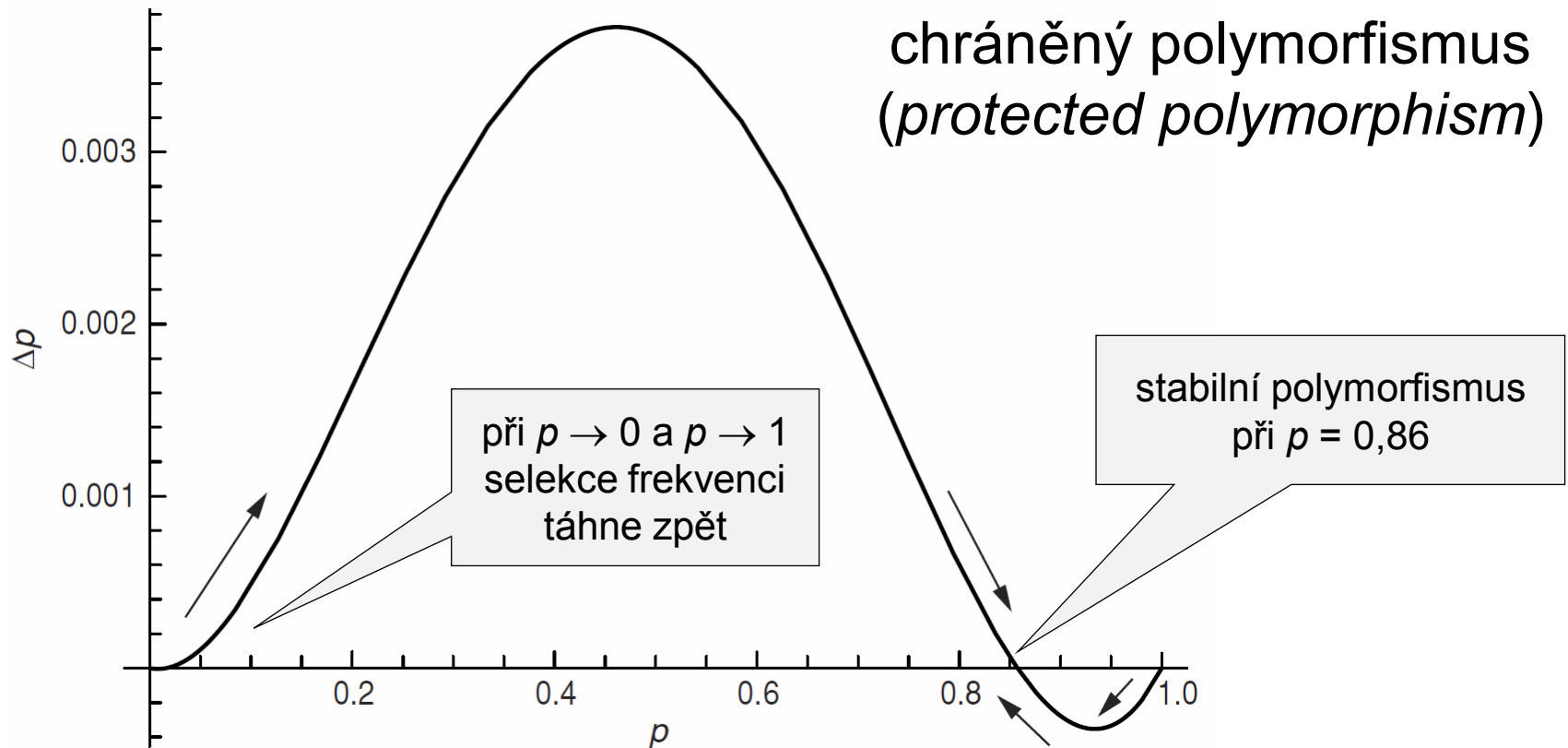
Habitat	<i>SS</i>	<i>SF</i>	<i>FF</i>
Exposed substrate in high intertidal zone	0.696	1	1.424
Exposed substrate in low intertidal zone	0.898	1	1.012
Under algal canopy in high intertidal zone	1.519	1	0.880
Under algal canopy in low intertidal zone	0.913	1	0.976

$c_1 = 0,15$
 $c_2 = 0,375$
 $c_3 = 0,325$
 $c_4 = 0,15$



Source: From Schmidt and Rand (2001).

Note: All values are measured relative to the fitness of the heterozygote.



⇒ kvalitativně odlišné výsledky při stejných hodnotách fitness jako funkce parametru c

Prostorová heterogenita v hrubém měřítku rozšiřuje podmínky pro chráněný polymorfismus

V homogenním prostředí a při konstantní fitness polymorfismus jen při superdominanci ($W_{AA} < W_{Aa} > W_{aa}$)

× v heterogenním prostředí ne, např.:

Table 14.1. Habitat-Specific Viability Estimates for *Mpi* Genotypes in Northern Acorn Barnacle

Habitat	<i>SS</i>	<i>SF</i>	<i>FF</i>
Exposed substrate in high intertidal zone	0.696	1	1.424
Exposed substrate in low intertidal zone	0.898	1	1.012
Under algal canopy in high intertidal zone	1.519	1	0.880
Under algal canopy in low intertidal zone	0.913	1	0.976

Source: From Schmidt and Rand (2001).

Note: All values are measured relative to the fitness of the heterozygote.

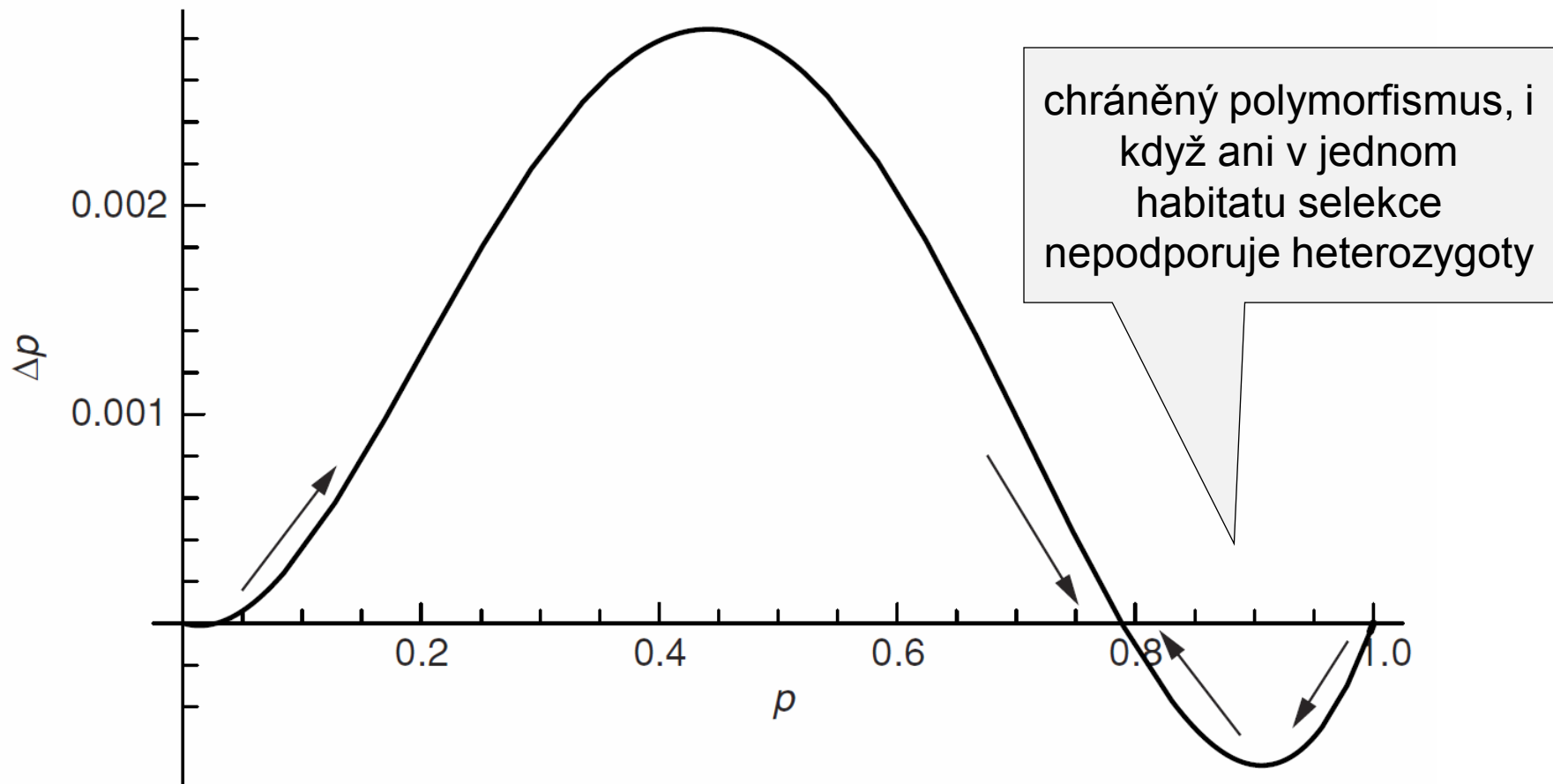
$$c_1 = 0,12$$

$$c_2 = 0,58$$

$$c_3 = 0,3$$

$$c_4 = 0$$

jediný habitat se superdominancí



tvrdá selekce:

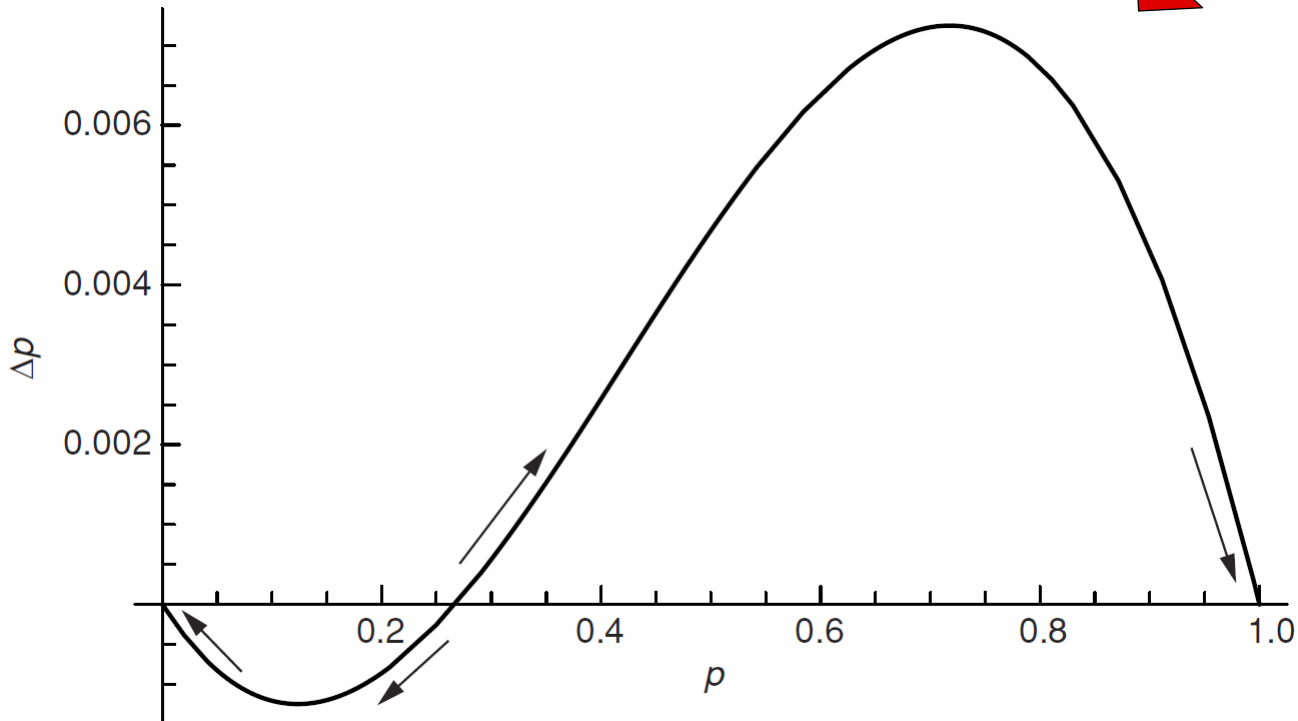
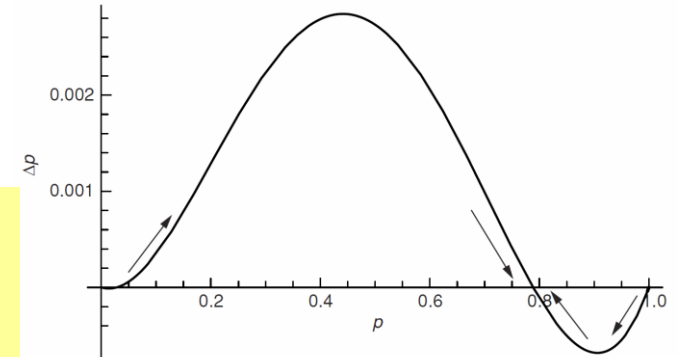
Table 14.1. Habitat-Specific Viability Estimates for *Mpi* Genotypes in Northern Acorn Barnacle

Habitat	<i>SS</i>	<i>SF</i>	<i>FF</i>
Exposed substrate in high intertidal zone	0.696	1	1.424
Exposed substrate in low intertidal zone	0.898	1	1.012
Under algal canopy in high intertidal zone	1.519	1	0.880
Under algal canopy in low intertidal zone	0.913	1	0.976

Source: From Schmidt and Rand (2001).

Note: All values are measured relative to the fitness of the heterozygote.

$c_1 = 0,12$
 $c_2 = 0,58$
 $c_3 = 0,3$
 $c_4 = 0$



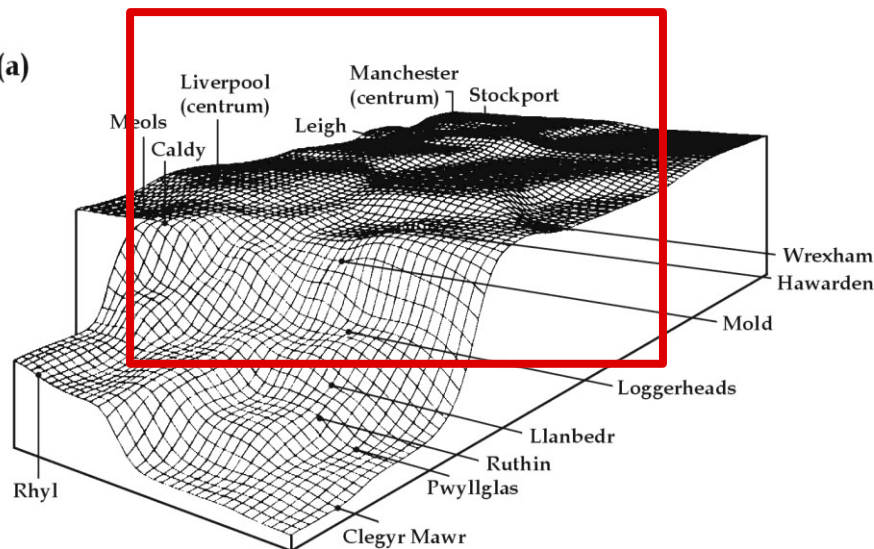
nestabilní i nechráněný polymorfismus

Při tvrdé selekci prostorová heterogenita v hrubém měřítku podmínky pro polymorfismus nerozšiřuje.

ALE:

Pokud tok genů velmi omezený, ekologický rozdíl mezi měkkou a tvrdou selekcí je z hlediska podmínek pro udržení polymorfismu chráněného prostorovou variabilitou v hrubém měřítku nepodstatný.

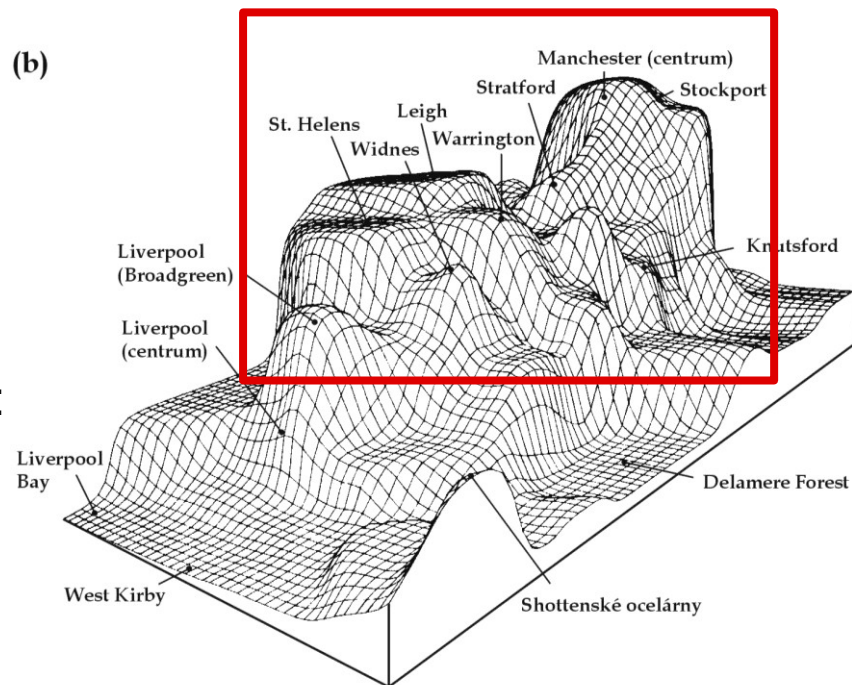
(a)



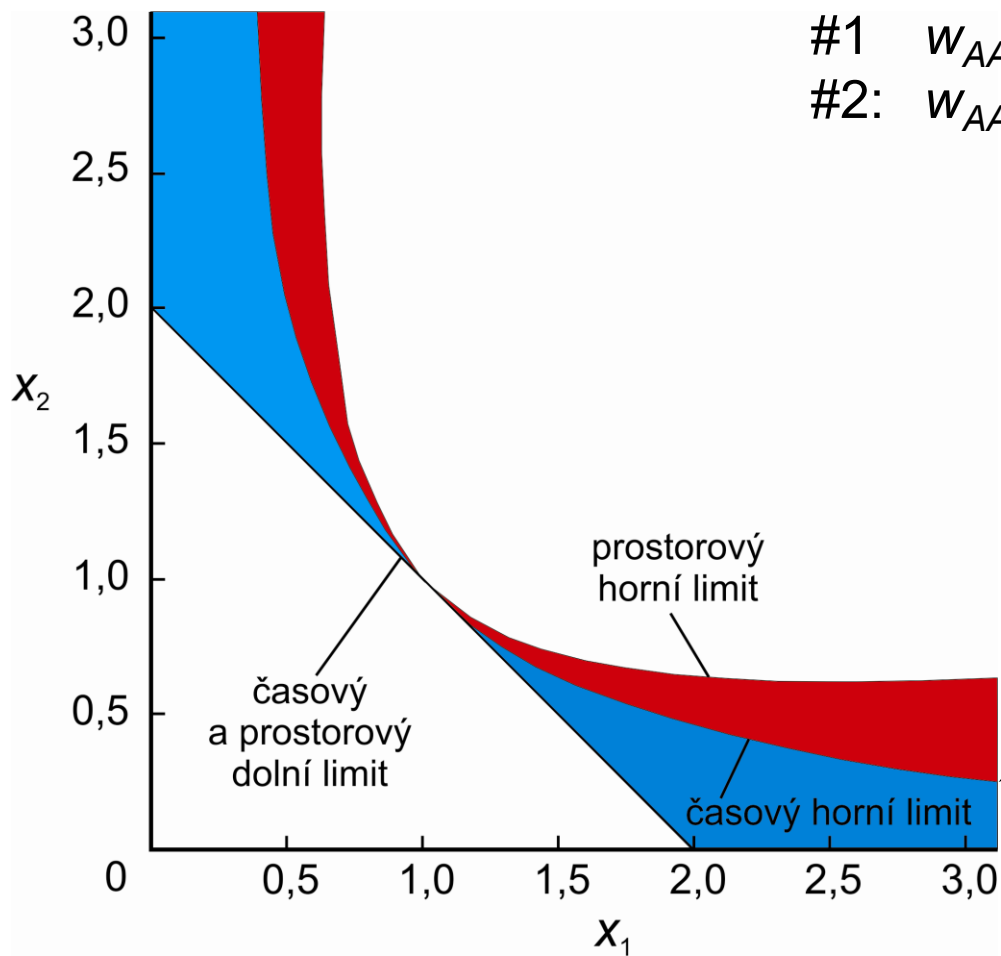
zejkovec dvojzubý (*Odontoptera bidentata*):
50 000–100 000/km²; migrace < 150 m

drsnokřídlec březový (*Biston betularia*):
~10/km²; migrace > 1 km

(b)



Vyšší tok genů → na lokální úrovni proměnlivost prostředí
v jemném měřítku; adaptace na prostředí měnící se
v hrubém měřítku na větší geografické škále.



#1 $W_{AA} = 1$ $W_{Aa} = 1$ $W_{aa} = X_1$
 #2: $W_{AA} = 1$ $W_{Aa} = 1$ $W_{aa} = X_2$

kombinace hodnot fitness poskytující stabilní polymorfismus je restriktivnější pro časovou než prostorovou variabilitu

2. Prostředí proměnlivé v jemném měřítku (*fine-grained environment*)

Př.: mutace $a \rightarrow A$; fitness nového genotypu $Aa = 1 + s$

σ_s^2 = heterogenita prostředí v jemném měřítku

Jaká je pravděpodobnost přežití alely A ?

pokud s malé, platí:

$$P(\text{přežití } A) = \frac{2s}{1 + s + \sigma_s^2}$$

při homogenitě prostředí
($\sigma_s^2 = 0$) $P \approx 2s$

2 mutace: $a \rightarrow A_1$ a $a \rightarrow A_2$; fitness obou heterozygotů stejná, tj. $1 + s$

pokud $\sigma_{s1}^2 < \sigma_{s2}^2 \Rightarrow A_1$ má větší šanci na přežití než A_2 !

\Rightarrow selekce podporuje mutace poskytující účinnější
pufrování proti fluktuacím v důsledku proměnlivosti
prostředí v jemném měřítku

Za určitých okolností může být pravděpodobnost přežití vyšší i u alely spojené s nižší fitness:

např. $s_1 > s_2$

→ za normálních okolností by šance na přežití byla vyšší pro A_1 než pro A_2

ale pokud platí

$$\sigma_{s_2}^2 < \frac{s_2 - s_1}{s_1} + \frac{s_2}{s_1} \sigma_{s_1}^2$$

bude mít vyšší pravděpodobnost přežití alela spojená s nižší fitness

tj. je-li genotyp A_1a dostatečně pufrován proti heterogenitě prostředí v jemném měřítku, bude mít mutace A_1 větší šanci v populaci přežít, než prospěšnější mutace A_2

Např.: $s_1 = 0,02$; $s_2 = 0,01$; $\sigma_{s_1}^2 = 4$; $\sigma_{s_2}^2 = 0,5$

žádná heterogenita: 2x vyšší šance přežití pro A_1 než pro A_2

heterogenita: $P(A_1) = 0,008$
 $P(A_2) = 0,013$ ← $P(\text{přežití } A) = \frac{2s}{1 + s + \sigma_s^2}$

⇒ A_2 je výhodnější

neutralita: $s_1 = s_2$, tj. stejná fitness × selekce bude preferovat genotyp lépe pufovaný proti fluktuacím fitness

Pokud konečná velikost populace:

$$P = \frac{1 - e^{-2s}}{1 - e^{-4Ns}}$$

výhodná alela;
viz Selekcce II.,
snímek # 20

při heterogenitě prostředí místo $s \rightarrow (s - \nu/2N)$,
kde selekční koeficient jedinice je náhodná veličina s průměrem s
a rozptylem $\nu/2N$

⇒ průměrná fitness v modelu proměnlivého prostředí
v jemném měřítku je při určení pravděpodobnosti fixace
výhodné alely snížena o $\nu/2N$

⇒ s klesající schopností pufování heterogenosti prostředí
v jemném měřítku (tj. s rostoucím rozptylem ν) klesá
pravděpodobnost fixace prospěšné alely

Pozor: je třeba mít na zřeteli, že znak, který je zde pufrován, je reprodukční zdatnost

pufrování fitness je ale většinou zprostředkováno jinými znaky, které jsou plastické a citlivé na heterogenitu prostředí v jemném měřítku

Např. člověk a jeho schopnost bránit se výkyvům teploty:

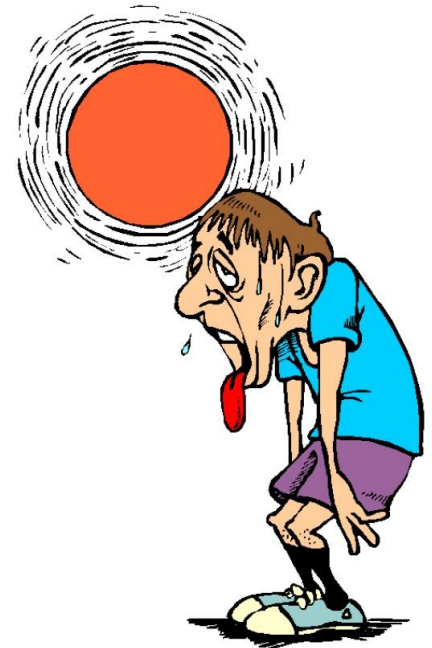
zvýšení teploty → dilatace povrchových cév

⇒ odvod tepla k povrchu těla

další zvýšení → pocení

⇒ další odvod tepla odpařováním

(delší expozice vysoké teploty ovšem může mít fatální následky)



snížení teploty → vazokonstrikce ⇒ snížení tepelných ztrát
další snížení → třes atd. ⇒ produkce tepla
(delší expozice chladu opět může mít fatální následky)



subsaharská Afrika: dřívější a intenzivnější pocení

Evropa: dřívější začátek třesu (teplota kůže $29,5\text{ }^{\circ}\text{C}$ × Afričané $28\text{ }^{\circ}\text{C}$)

⇒ dřívější nástup pufrujícího mechanismu v populaci, která má historicky větší zkušenost

Kromě intenzity má vliv i délka působení prostředí

→ při dlouhodobějším působení jiné mechanismy
(např. zvýšení počtu potních žláz)

Prostředí proměnlivé **v hrubém měřítku a měkká selekce** budou v populaci udržovat polymorfismus s vyšší pravděpodobností než proměnlivost v jemném měřítku a tvrdá selekce.