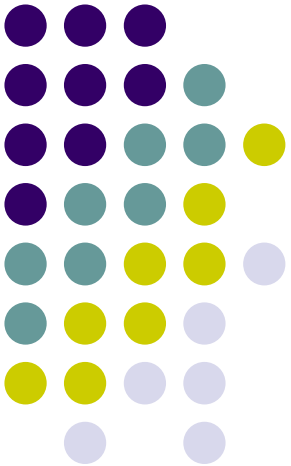
Genetické algoritmy

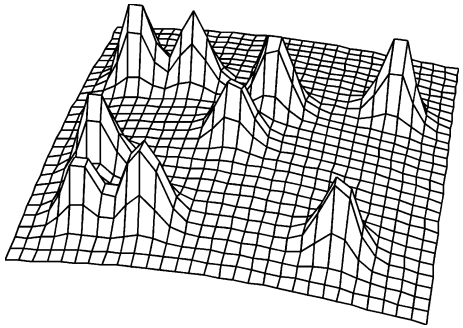
Spracované podľa knihy V. Kvasničku a J. Pospíchala

- Darvinovská evolučná teória je v súčasnosti charakterizovaná ako ***univerzálny algoritmus*** s platnosťou nielen v biológii, ale aj v iných oblastiach ľudského poznania,
- hlavne tam, kde sme schopní vyabstrahovať **informačné entity - replikátory**, ktoré majú schopnosť reprodukovať sa a medzi ktorými prebieha prirodzený výber.

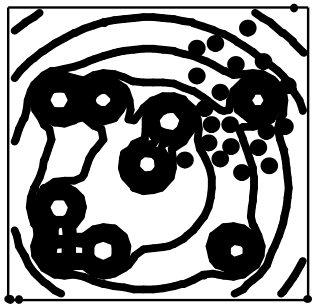


Genetické algoritmy

Genetik S. Wright v 30. rokoch minulého storočia charakterizoval ***evolúciu*** *ako* ***optimalizáciu na povrchu fitnes funkcie*** (fitness landscape), kde sa hľadá genotyp odpovedajúci globálnemu maximu.

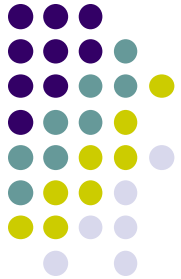


A



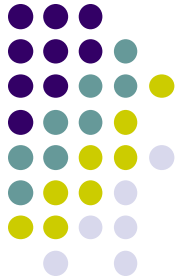
B

Čo vieme o genetike



- Genetika je jednou z najdôležitejších (respektíve priamo najdôležitejšia) teoretických vied z hľadiska popisu akejkoľvek živej sústavy.
- V genetickej informácii je počiatok každého súčasného živého organizmu.
- Genetická informácia
  - určuje budúcu anatomickú stavbu organizmu,
  - určuje aké látky budú súčasťou biochemických a fyziologických procesov v organizme a
  - je nezameniteľnou súčasťou pohlavného aj nepohlavného rozmnožovania.

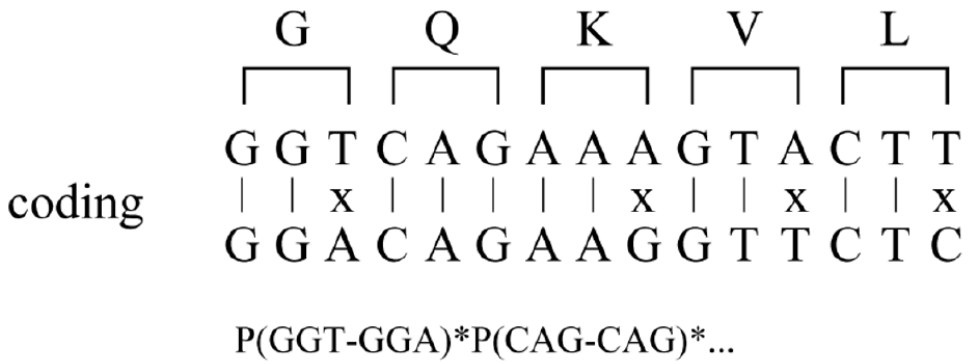
ZÁKLADNÉ POJMY



- **genetika** - veda o dedičnosti a premenlivosti
- **heredita** - /dedičnosť/, schopnosť organizmov odovzdávať vlohy pre utvorenie konkrétneho znaku
- **variabilita** - /premenlivosť/, schopnosť jedincov v rámci toho istého druhu líšiť sa od seba navzájom
- **gén** - úsek molekuly DNA, ktorý nesie úplnú genetickú informáciu pre vytvorenie určitej vlastnosti
  - základná funkčná jednotka dedičnosti

Základné pojmy

- **alela** - konkrétna forma génu
- rôzne alely podmieňujú rozdielny prejav znaku
- **Gén je zodpovedný za farbu očí. Alela modrá, zelená.... farba.**
- **S DNA súvisia RNA sekvencie. Ich popis sa robí pomocou abecedy 4 znakov** A = {A, T, C, G}.



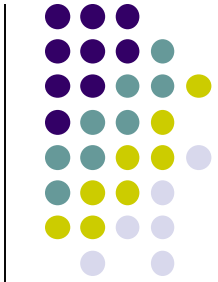
P(GGT-GGA)*P(CAG-CAG)*...

Základné pojmy

- **genotyp** - súbor génov v organizme, tak ako sa prejavujú v konkrétnych alelách
- **lokus** - konkrétne miesto génu na chromozóme

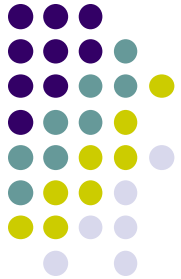
- **chromozómy** - sú prevažne uložené v jadre – **jadrová dedičnosť**
- **fenotyp** - súbor všetkých znakov v organizme, tak ako sa prejavujú v konkrétnych kvalitatívnych formách /alelách/ a kvantitatívnych stupňoch

Genetické algoritmy



- V informatike našla táto zaujímavá idea svoj odraz už pred viac ako 30 rokmi, keď **John Holland** vynašiel genetické algoritmy, ktoré je možno chápať ako algoritmy darvinovskej evolúcie a ktoré sa stali v súčasnosti rozvíjajúcou sa oblasťou informatiky.

Darvinovské systémy a univerzálny darvinizmus

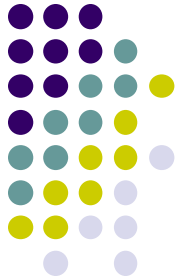


Najvšeobecnejšia formulácia základnej idey univerzálneho darvinizmu je pomocou koncepcie ***darvinovského systému* (DS)** - **dva postuláty**:

(1) *DS sa skladá z* ***populácie replikátorov*** – jedincov/ objektov, ktoré za určitých vhodných podmienok sú schopné ***replikácie*** - rozmnožovania. Replikačný proces spočíva v „kopírovaní“ jedincov do populácie, pričom toto „kopírovanie“ sa uskutočňuje s *určitými malými chybami*.

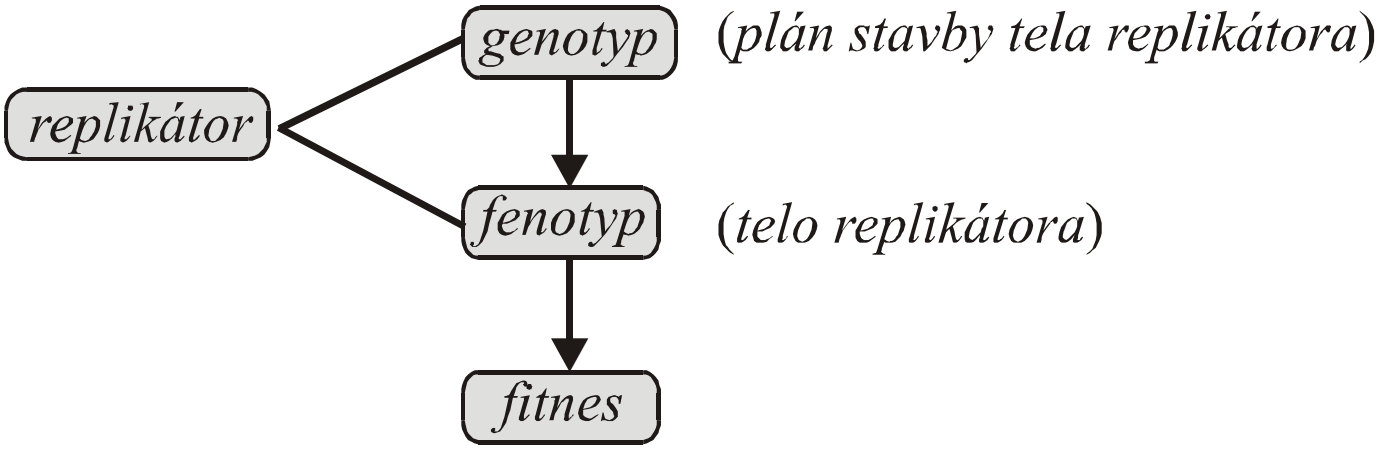
(2) Každý replikátor populácie je ohodnotený ***fitnes (silou) hodnotou***, ktorá vyjadruje schopnosť replikátora prežiť a úspešne vstupovať do replikačného procesu.

Darvinovské systémy a univerzálny darvinizmus



Prírodzený výber v darvinovskom systéme spočíva v tom, že jedinci populácie nie sú vyberaní do replikačného procesu náhodne, ale s *pravdepodobnosťou úmernou ich fitnes* (hovoríme, že výber sa deje kvázi náhodne).

Replikátor - informácia, ktorá kóduje „telo“ replikátora. Rozlišujeme dve rôzne špecifikácie replikátora, jeho *genotyp* a *fenotyp*.



- ***Fenotyp*** – organizmus replikátora – nosič (vehikel) genotypu, ktorý umožňuje jeho replikáciu.
- ***Proces replikácie*** sa chápe ako kopírovanie genotypu, pričom tento proces kopírovania je „fyzicky“ uskutočnený fenotypom replikátora.

Existujú darvinovské systémy, kde *odlíšenie fenotypu od genotypu neplatí*, **(biologické a počítačové vírusy, ktoré k vlastnej replikácii využívajú systémy, v ktorých parazitujú).**

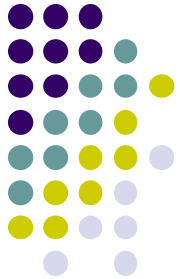
- Replikátor je reprezentovaný svojím *genotypom* ***x***, ktorý, obsahuje informáciu o stavbe replikátora. *Populácia* replikátorov je multimnožina genotypov

$$P = \left\{ \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p \right\}$$

- Vzájomný vzťah medzi triádou „genotyp – fenotyp – fitnes“ je reprezentovaný postupnosťou dvoch zobrazení

$$G \xrightarrow{\text{fenotyp}} F \xrightarrow{\text{fitnes}} [0, \infty)$$

Replikačný proces budeme rozlišovať ako unárny (asexuálny) a binárny (sexuálny).



- Rodičia (rodič) sú kvázi náhodne vybraní z populácie v závislosti od ich fitnes hodnoty (replikátory s väčším fitnes s väčšou pravdepodobnosťou vstupujú do replikácie) a produkujú nové replikátory - potomkov.
- Budeme rozlišovať tieto tri zložky replikačného procesu
  - ***selekcia*** rodičov,
  - ***replikácia*** rodičov, a
  - ***návrat*** potomkov do populácie

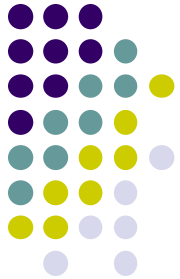
$$\boldsymbol{x}_1^{old} = O_{select}\left(P\right), \quad \boldsymbol{x}_2^{old} = O_{select}\left(P\right)$$

$$\left(\boldsymbol{x}_1^{new}, \boldsymbol{x}_2^{new}\right) = O_{repli}\left(\boldsymbol{x}_1^{old}, \boldsymbol{x}_2^{old}\right)$$

- V unárnej (asexuálnej) replikácii sa na tvorbe potomkov podieľa len jeden replikátor – rodič,

$$\boldsymbol{x}^{old} = O_{select}\left(P\right) \quad \text{a} \quad \boldsymbol{x}^{new} = O_{repro}\left(\boldsymbol{x}^{old}\right)$$

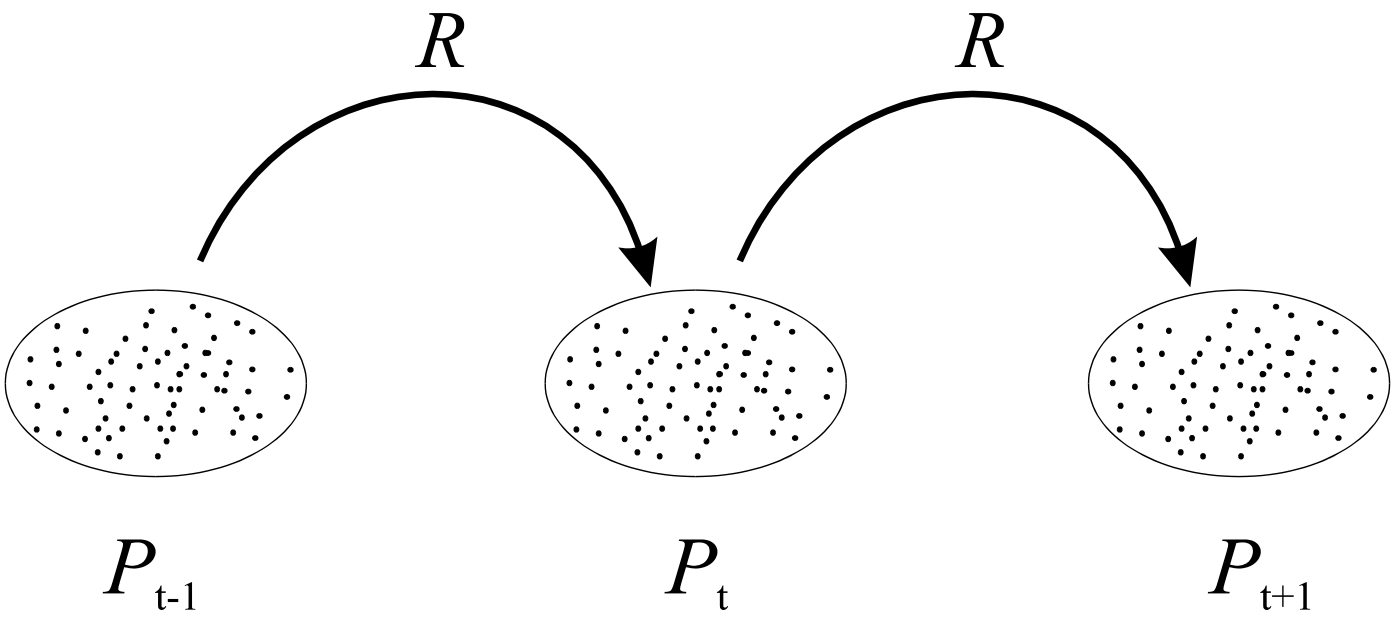
Pseudokód algoritmu univerzálnej Darwinovej evolúcie



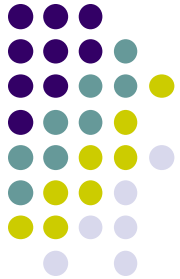
P:= náhodne vygenerovaná populácia replikátorov;
t:=0;
pokiaľ t<tmax **urob**
{ t:=t+1;
Q:=∅;
pokiaľ |Q|<|P| **urob**
{
$x1:=O_{select}(P)$;
$x2:=O_{select}(P)$;
$(x1', x2'):=O_{repli}(x1, x2)$;
$Q:=Q\cup\{x1', x2'\}$;
}
P:=Q;
}

Darwinova evolúcia môže byť interpretovaná ako *rekurentný proces*, v ktorom nasledujúca populácia je vytvorená reprodukciou predchádzajúcej populácie

$$P_{t+1} = R(P_t)$$



Genetické algoritmy



- Genetické algoritmy - John Holland začiatkom 70-tých rokov minulého storočia.

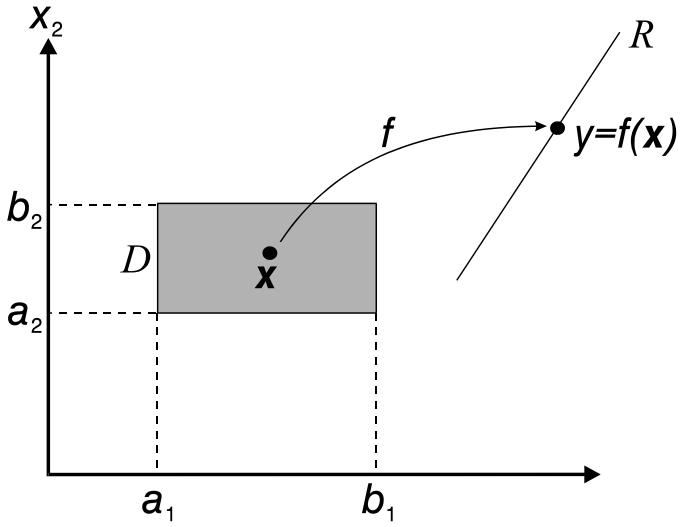
- Po určitej nábehovej *10-ročnej* perióde rozpakov a mlčania v komunite informatikov sa stali jednou z rozvíjajúcich sa oblastí informatiky a umelej inteligencie.
- Spolu s neurónovými sieťami tvoria jadro oblasti nazývanej počítačová inteligencia, ktorá je schopná riešiť už praktické problémy z informačných technológií, ktoré majú vysoký stupeň „inteligentnosti“.

Optimalizačný problém

Nech funkcia $f: D \rightarrow R$

$$D = \prod_{i=1}^{n} a_i, b_i = a_1, b_1 \times a_2, b_2 \times \ldots \times a_n, b_n$$

zobrazuje n - rozmernú kocku D (karteziánsky súčin uzavretých intervalov $[a_i, b_i]$) na reálne čísla $y \in R$



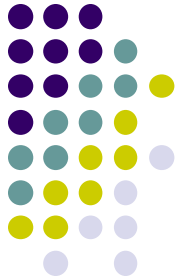
Nech táto funkcia spĺňa dve podmienky:

(1) Existuje taký algoritmus, ktorý funkciu *f* "vypočíta dostatočne rýchlo" s požadovanou presnosťou pre každé $\boldsymbol{x} \in D$ (hovoríme, že funkcia f je *dobre vypočítateľná*).

(2) Pre každú dvojicu lokálnych miním $\boldsymbol{x}1, \boldsymbol{x}2 \in D$ vzdialenosť $|\boldsymbol{x}1-\boldsymbol{x}2|$ je väčšia ako dané kladné číslo $\delta>0$, $|\boldsymbol{x}1-\boldsymbol{x}2|>\delta$.

Podmienka ohraničuje zhora počet lokálnych miním funkcie *f*, ktoré sa vyskytujú na kocke *D.* Nie je možné, aby sa v ľubovoľnom okolí minima funkcie vyskytovalo iné minimum, pre určité malé okolie minima funkcie vyššie uvedená podmienka $|\boldsymbol{x}1-\boldsymbol{x}2|>\delta$ by prestala platiť.

Podmienka automaticky vylučuje z triedy prípustných funkcií tie funkcie, ktoré sú "fraktálového" typu, t.j. v každom okolí nejakého minima sa nachádza aspoň jedno iné minimum.



Globálne maximum funkcie *f* na kocke *D* je určené vzťahom

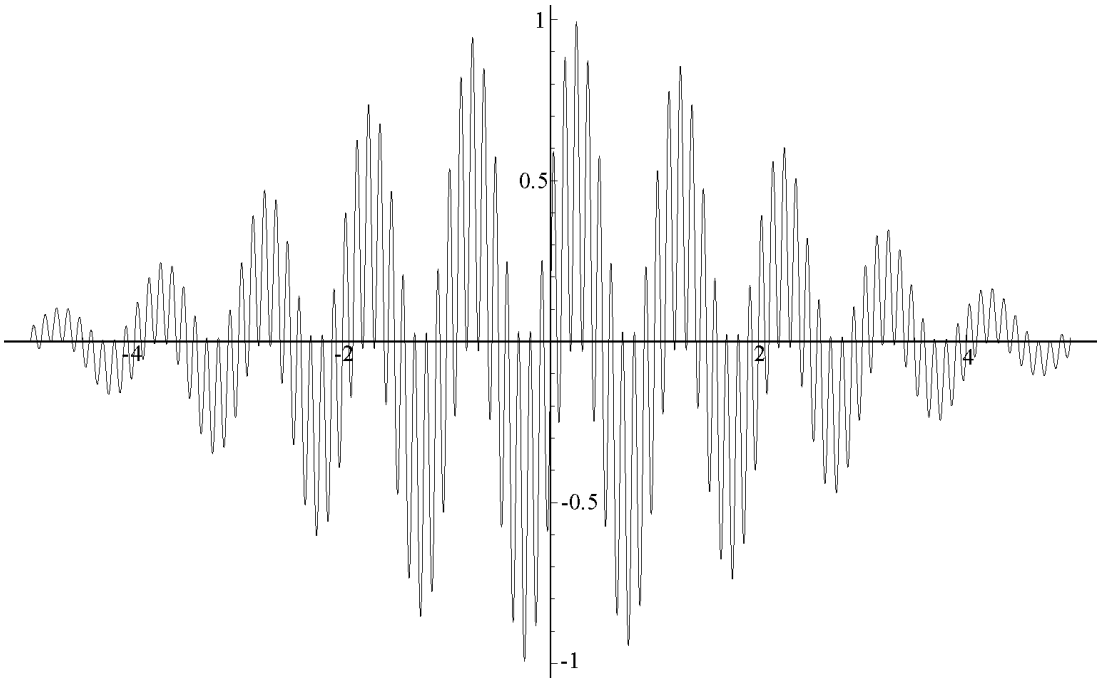
$$x_{opt} = \arg\max_{\mathbf{x} \in D} f(\mathbf{x})$$

je teda určené ako argument, v ktorom funkcia nadobúda maximum na D.

Príklad:

$$F(x) = e^{-0.1x^2}\ sin(10\pi x)cos(8\pi x) \quad \forall x \in [-5,5]$$

Globálne maximum funkcie, jeho presná hodnota je získaná použitím Newtonovej metódy: $f(x_{opt}) = 0.99377, x_{opt} = 0.249969$



Základný problém použitia genetických algoritmov je binárna reprezentácia reálnych čísel.

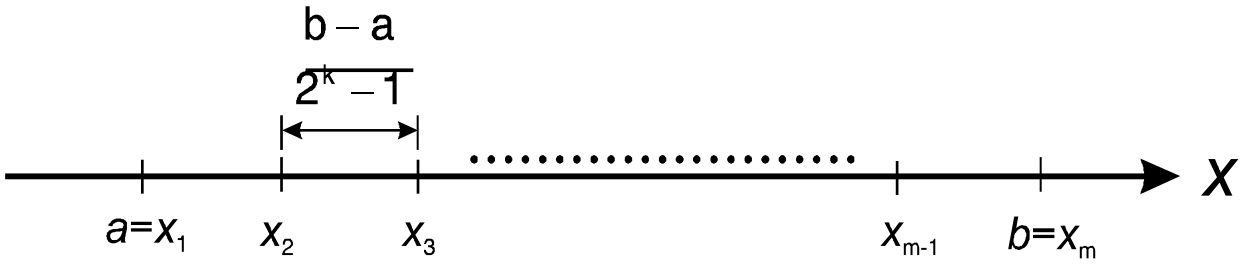
- Nech binárny vektor α dĺžky k je interpretovaný ako celé číslo

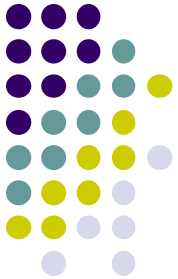
$$int(\alpha) = \sum_{i=1}^{k} \alpha_i 2^{k-i} = \alpha_1 2^{k-1} + \alpha_2 2^{k-2} + \ldots + \alpha_{k-1} 2 + \alpha_k$$

- K tomuto celému číslu jednoduchým spôsobom priradíme reálne číslo, ktoré môže byť chápané ako aproximácia reálneho čísla $x \in [a, b]$

$$x \approx real(\alpha) = a + \frac{b-a}{2^k - 1} int(\alpha)$$

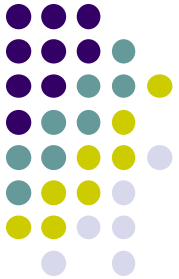
- Táto konštrukcia racionálneho čísla $a \leq real(\alpha) \leq b$ z binárneho reťazca α dĺžky k sa formálne interpretuje ako "transformácia" binárnej reprezentácie na "reálnu" reprezentáciu,
- zostrojené racionálne číslo real(α) aproximuje požadované reálne číslo x s presnosťou $(b-a)/(2k-1)$. Interval $[a, b]$ obsahuje $m=2k$ bodov $x_1=a$, $x_2=a+(b-a)/(2k-1)$, ..., $x_i = a+(i-1)(b-a)/(2k-1)$, ..., $x_n=b$,





Tabuľka 2.1. **Reprezentácia čísel**

No.	α	int(α)	real(α)	(Gray)
1	000	0	0	000
2	001	1	1/7	001
3	010	2	2/7	011
4	011	3	3/7	010
5	100	4	4/7	110
6	101	5	5/7	111
7	110	6	6/7	101
8	111	7	1	100



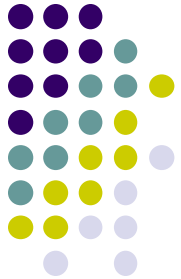
- Inverzná transformácia má tvar

$$\mathrm{int}(\boldsymbol{\alpha}) = \left\lceil \frac{x-a}{b-a}\left(2^{k}-1\right) \right\rceil$$

- Prechod od binárneho vektora $\boldsymbol{\alpha}=(\alpha 1,\alpha 2, \ldots, \alpha kn)\in\{0,1\}$ *kn* k spojitému vektoru $\boldsymbol{x}=(x1,x2,\ldots, xn)\in D$ sa môže formálne chápať ako transformácia

$$\Gamma : \{0,1\}^{kn} \to D \qquad \boldsymbol{x} = \Gamma(\boldsymbol{\alpha})$$

- ktorá zobrazuje množinu binárnych vektorov dĺžky *kn* na body - *n*-tice reálnych čísel *z* kocky *D.* Ináč povedané, konečná množina (2*kn*) binárnych vektorov dĺžky *kn* je reprezentovaná pomocou zobrazenia Γ bodmi, ktoré môžu byť v oblasti *D* usporiadané do ortogonálnej mriežky.



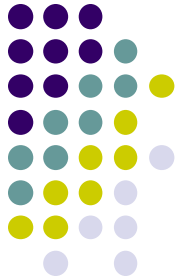
Nech *F*(*x*) je účelová funkcia, definovaná na intervale [*a*, *b*], na ktorom budeme hľadať jej maximálnu hodnotu

$$F\left(x_{opt}\right) = \max_{x \in [a,b]} F\left(x\right)$$

Predpokladajme, že reálne čísla z intervalu [*a*, *b*] sú aproximované binárnymi reťazcami dĺžky *k*, potom spojitý optimalizačný problém je prevedený na diskrétny optimalizačný problém nad binárnymi reťazcami dĺžky *k*

$$F\left(\alpha_{opt}\right) = \max_{\alpha \in \{0,1\}^k} F\left(\alpha\right)$$

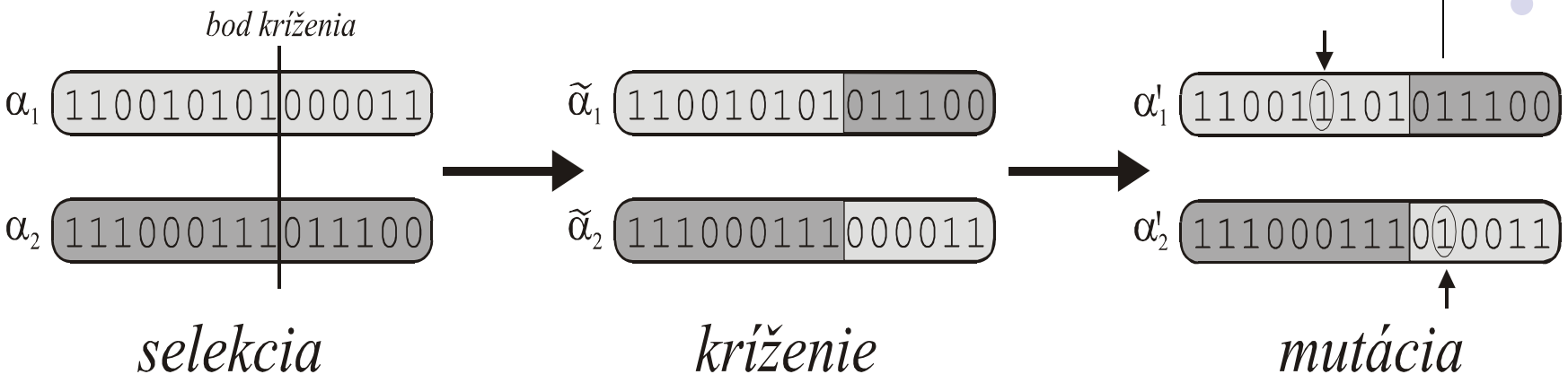
Ohodnotenie replikátorov veličinou fitnes v genetickom algoritme sa robí pomocou funkcie, ktorú chceme optimalizovať.



- Majme dva replikátory a1 a a2, ich ohodnotenie fitnes musí vyhovovať podmienke , kde *f*(a) je fitnes priradené replikátoru – reťazcu a.
- Všeobecný operátor replikácie O_{repli}, definovaný v rámci univerzálneho darvinizmu, má v prípade genetických algoritmov dve časti: ***kríženie a mutáciu*.**

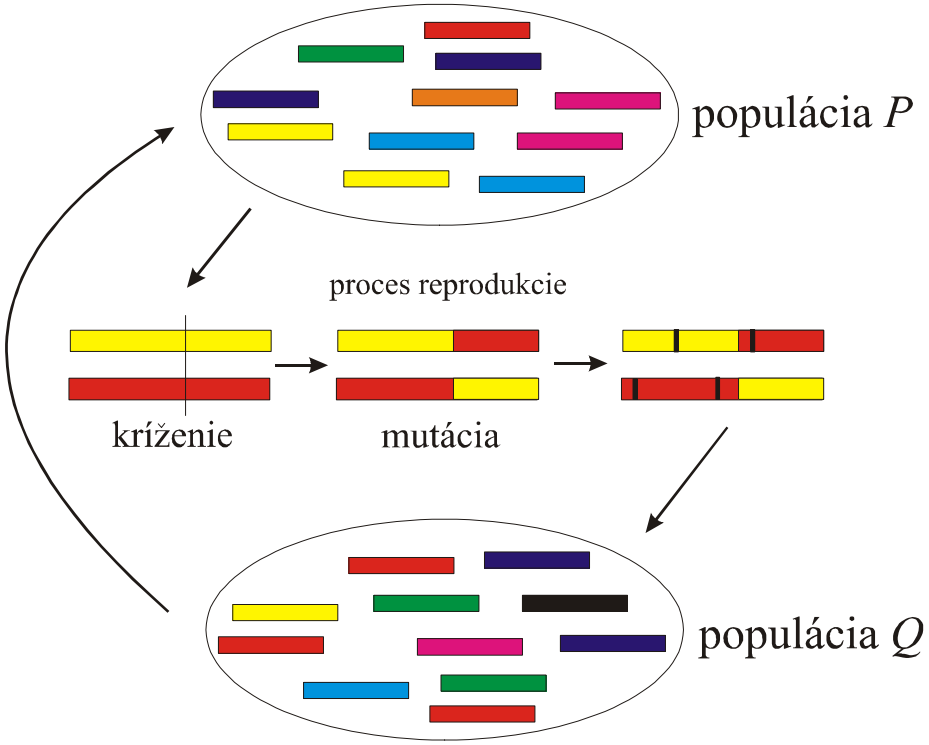
selekcia	$\alpha_1 = O_{select}(P)$ a $\alpha_2 = O_{select}(P)$
kríženie	$(\tilde{\alpha}_1, \tilde{\alpha}_2) = O_{cross}(\alpha_1, \alpha_2)$
mutácia	$(\tilde{\alpha}_1, \tilde{\alpha}_2) = O_{cross}(\alpha_1, \alpha_2)$a $(\tilde{\alpha}_1, \tilde{\alpha}_2) = O_{cross}(\alpha_1, \alpha_2)$

Genetické algoritmy



• V prvej etape sú vybrané dva replikátory – rodičia s pravdepodobnosťou úmernou ich fitnes.
• V druhej etape, pri krížení, si replikátory pri kopírovaní prehodia určité časti svojich genotypov – binárnych reťazcov.
• V tretej etape, pri mutácii, sa v nových replikátoroch – potomkoch – v niektorých polohách reťazca s malou pravdepodobnosťou zmenia hodnoty genotypu.

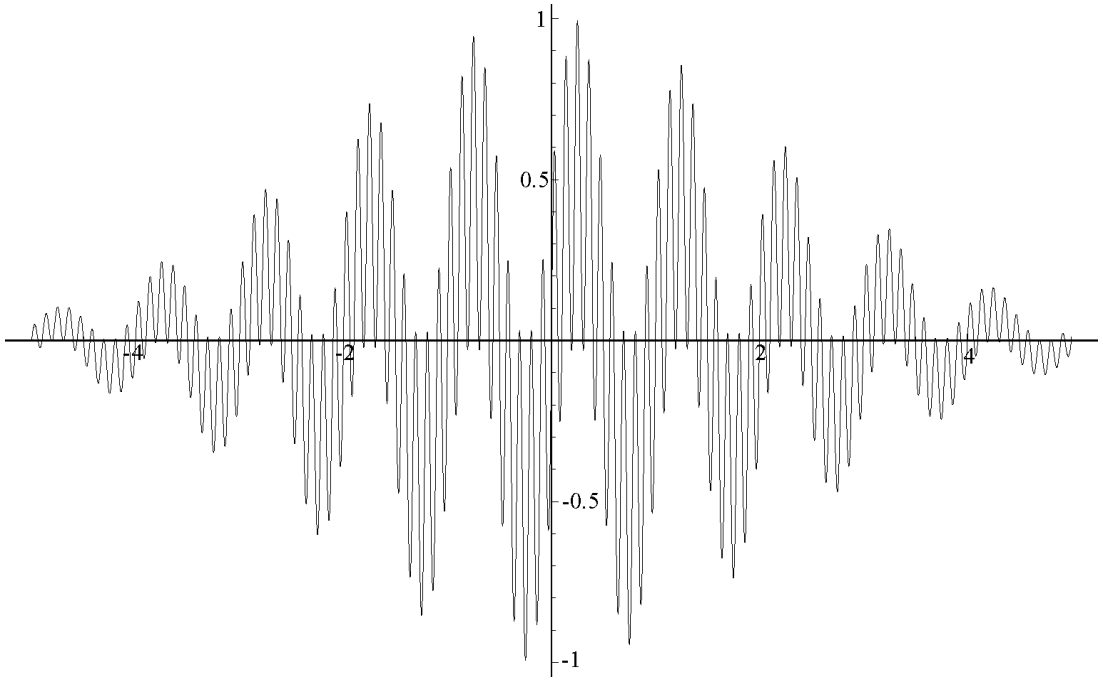
Diagramatická vizualizácia GA



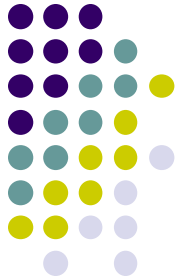
Aplikácia genetického algoritmu ako globálneho optimalizátora

$$F(x) = e^{-0.1x^2}\ sin(10\pi x)cos(8\pi x) \quad \forall x \in [-5,5]$$

Globálne maximum funkcie, jeho presná hodnota je získaná použitím Newtonovej metódy: $f(x_{opt}) = 0.99377, x_{opt} = 0.249969$



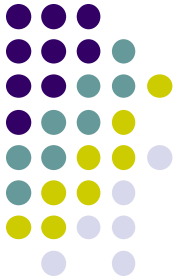
V simulačných výpočtoch predpokladali (Kvasnička, Pospíchal):



- Populácia obsahuje 200 binárnych replikátorov dĺžky *k*=30. Dá sa ukázať, že takto zvolená binárna reprezentácia špecifikuje reálne číslo z intervalu [-5, 5] s presnosťou na 7 dekadických miest za desatinnou bodkou.

- α = 0000010001|0010101000|0011100111
- Int(α) = 1*2^24+1*2^20+ 1*2^17+1*2^15+1*2^13+231=
- = (2^11+2^7+2^4+2^2+1)*2^13+231= 2193*8192+231=
- = 17965056 + 231 = 17965287
- a= -5, b= 5
- Real(α) = -5 + 179652870/(2^30-1) = -5+179652870/1073741824

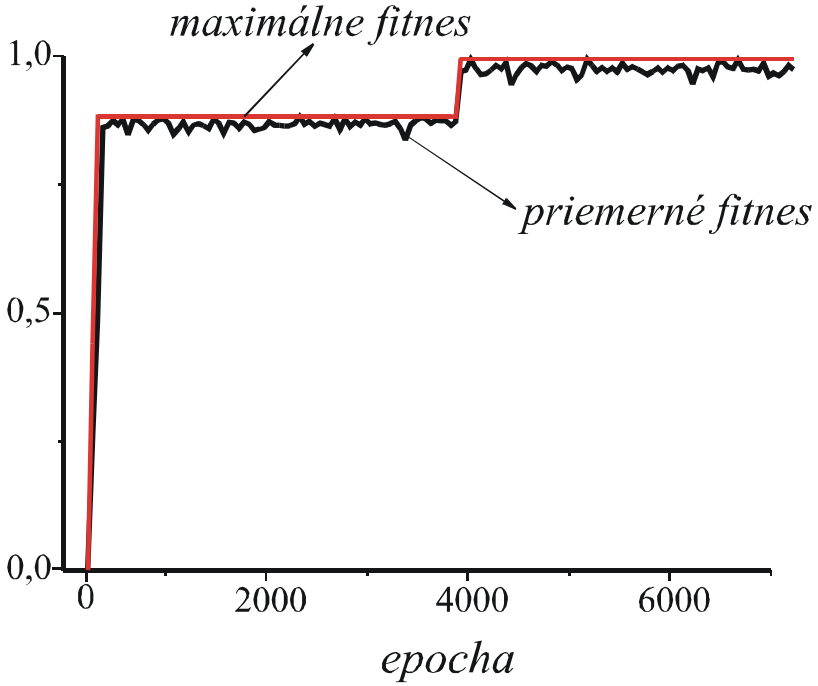
V simulačných výpočtoch predpokladali (Kvasnička, Pospíchal):



- Populácia obsahuje 200 binárnych replikátorov dĺžky k=30. Dá sa ukázať, že takto zvolená binárna reprezentácia špecifikuje reálne číslo z intervalu [-5, 5] s presnosťou na 7 dekadických miest za desatinnou bodkou.
- Fitnes replikátorov bude stotožnený priamo s funkčnou hodnotou účelovej funkcie .
- Výskyt mutácií v replikačnom procese je určený pravdepodobnosťou jednobodovej mutácie P_{mut}=0.001, ktorá sa interpretuje tak, že idúc po reťazci replikátora, každá jeho binárna hodnota je zmenená (zmutovaná) s pravdepodobnosťou P_{mut}.
- Populácia bola inicializovaná replikátormi – binárnymi reťazcami obsahujúcimi len nuly.

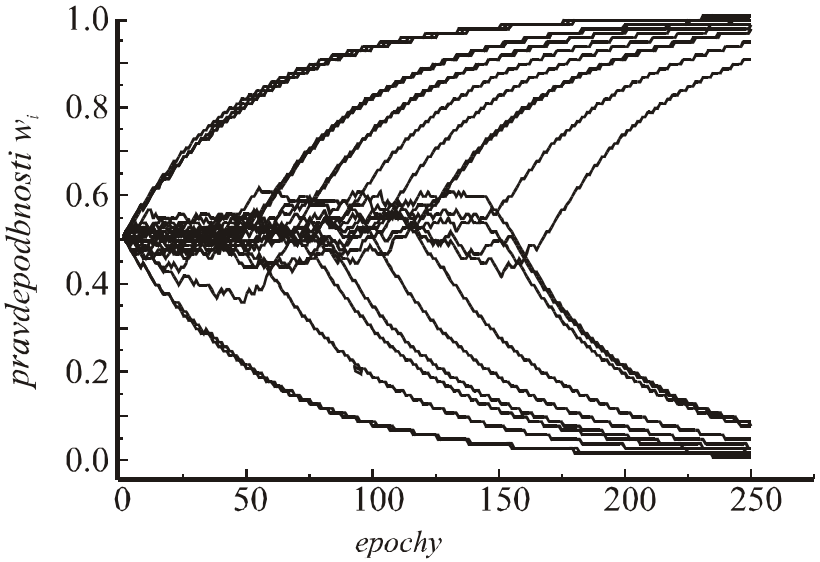
Priebeh priemerného a maximálneho fitnes v priebehu genetického algoritmu aplikovaného k hľadaniu globálneho maxima.
Priebeh fitness tvorí schodovú funkciu, na ktorej existujú pomerne dlhé etapy neutrality, v ktorých sa čaká na vhodnú mutáciu, ktorá zvýši funkčnú hodnotu.

Približne od epochy *t*=4000 populácia väčšinou obsahovala replikátory, ktoré odpovedali optimálnemu riešeniu .



Problém zastavenia genetického algoritmu

- Obvykle je inicializovaný populáciou replikátorov, ktoré sú náhodne generované.
- Evolúcia je zložená z opakovanej obnovy populácie pomocou rekurzie , kde populácia P_t je nahradená novou populáciou P_{t+1} pomocou procesu reprodukcie R.
- Proces prírodného výberu ponechá v populácii len tie replikátory, ktorých ohodnotenie je veľmi blízke k aktuálnemu maximálnemu fitnes (alebo je s ním totožné).



Priebeh pravdepodobnosti w_i pre evolúciu populácie replikátorov v genetickom algoritme. Populácia $P0$ bola inicializovaná náhodne generovanými binárnymi replikátormi, v priebehu evolúcie jednotlivé pravdepodobnosti sa asymptoticky blížia buď k jednotkovej hodnote, alebo k nulovej hodnote.

Úloha

Výpočet minimálnej hodnoty funkcie

z=1+(x-1)*(x-1)+y*y na intervale **<0, 2>x<0, 2>**

pri nasledujúcich predpokladoch: tmax=4, n=2, k=5, prvé tri náhodne vygenerované replikátory sú: 1010100010, 1100000110, 1010101010, a ďalšie podľa potreby vygenerujte sami. Pracujte s populáciou veľkosti 3.

Iné potrebné parametre zvoľte sami.

Záver

- Evolučné algoritmy patria medzi základné prostriedky modernej numerickej matematiky pre riešenie zložitých optimalizačných problémov.
- Používajú sa vtedy, ak hľadáme také globálne minimum, ktoré je obklopené množstvom lokálnych miním.
- Skutočnosť, že sa dajú takto použiť, je prekvapujúca (podobne ako pre neurónové siete vlastnosť, že sú univerzálnym aproximátorom funkcií),
- adaptáciou a modifikáciou všeobecných predstáv o Darwinovej evolučnej teórii sme dostali univerzálnu numerickú optimalizačnú metódu.