Algoritmi genetici: evolutie in genetica si informatica

Lorentz JÄNTSCHI

Catre genetica moderna

- Carl LINNAEUS (Linnaei, 1735) – prima clasificare a vieţuitoarelor după clase, ordine, genuri şi specii (şi denumirile acestora în Latină)
- Jean-Baptiste LAMARCK (Lamarck, 1809) – studii sistematice cu privire la trăsăturile şi caracterele plantelor; moştenirea uşoară
- Charles Robert DARWIN (Darwin, 1859) – teoria evoluției; supraviețuirea celui mai tare
- Gregor Johann MENDEL (Mendel, 1866) – moştenirea caracterelor prin încrucişare (studii pe mazăre)
- Friedrich Leopold August WEISMANN (Weismann, 1893) - moştenirea dură
- Thomas Hunt MORGAN (Morgan şi alții, 1915) - teoria cromozomială a moştenirii
- Sir Ronald Aylmer FISHER (Fisher 1918, 1922, şi 1954) - demonstrarea teoriei evoluției, selecției naturale, legilor lui Mendel, şi teoriei cromozomiale

Catre programarea evolutiva

- Primele simulări ale evoluției se regăsesc în studiile lui Nils Aall BARRICELLI (Barricelli, 1954)
- Alex FRASER (1923-2002) a publicat o serie de lucrări despre simularea selecției artificiale a organismelor cu locuşi multipli (locus: poziția pe care o anumită genă o ocupă pe un cromozom) ce controlează o trăsătură măsurabilă
- Simulările lui Fraser (Fraser, 1957-1970) includ toate elementele esențiale ale algoritmilor genetici moderni

Solutia informatica a complexitatii

- Complexitate exponențială: cel mai bun algoritm rezolvă problema într-un timp de execuție ce creşte exponențial cu volumul datelor de intrare
- Euristici: set de reguli ce rezolva o problema bazat pe bunul simt în ceea ce priveşte soluția aşteptată prin evitarea erorilor grosolane; nu produc totdeauna soluția cu exactitate şi respectiv nu sunt capabili să producă o soluție pentru orice valori de intrare
- Meta-euristici: euristici aplicabili la o mare varietate de probleme dificile

Algoritmii genetici: ultima generatie de meta-euristici

- Căutarea tabu (TS - Tabu Search; Glower, 1977 şi 1986; Glover şi alții, 1992)
- Călirea simulată (SA - Simulated Annealing; van Laarhoven and Aarts, 1987; Davis, 1987)
- Algoritmi genetici (în engleză GA - Genetic Algorithm): studii de amploare - după 1970 (Bosworth şi alții, 1972; Holland, 1975); re-inventați dupa 1991 (Davis, 1991, Holland, 1992)

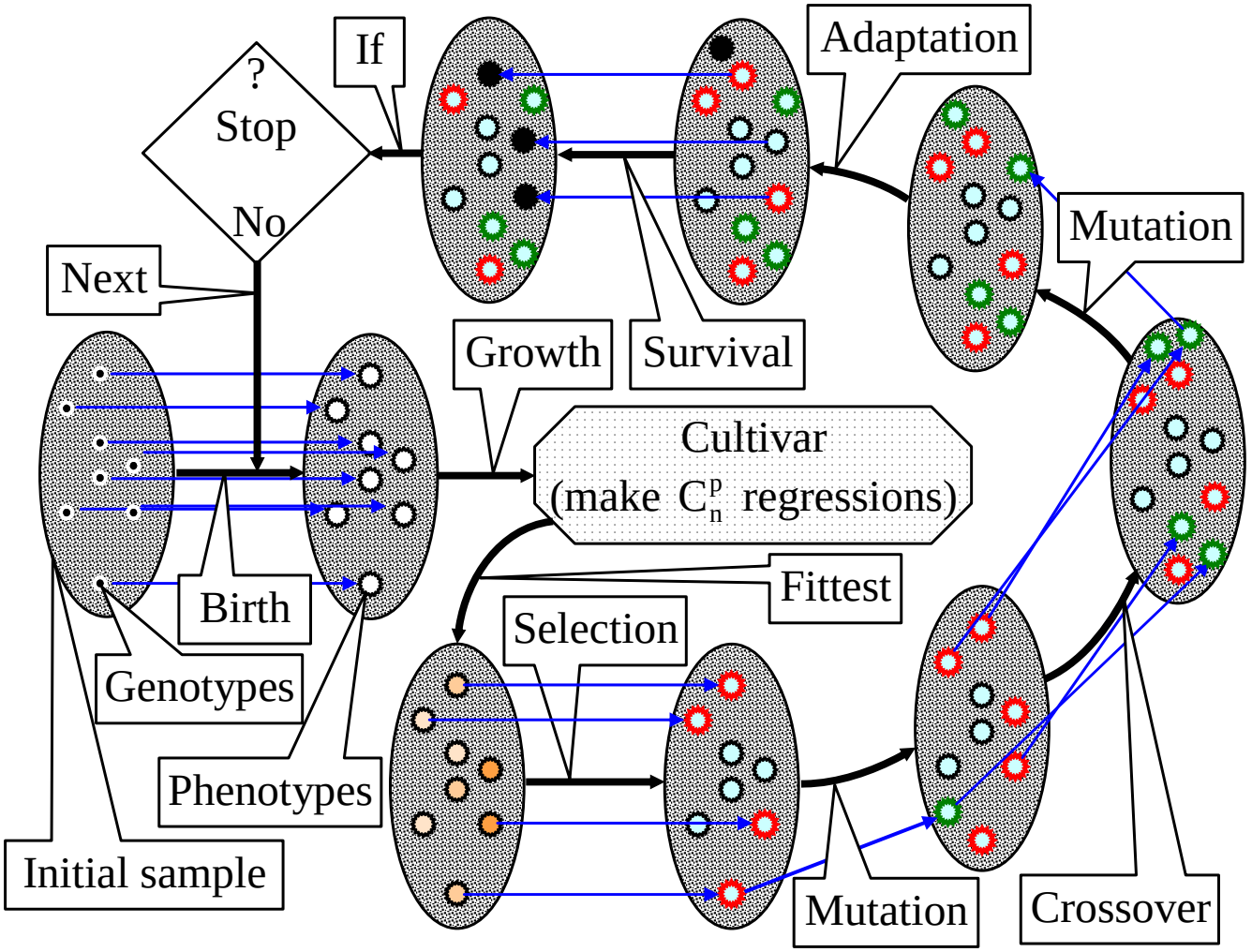
Probleme complexe

- Teorema *inexistenței mesei pe gratis* (NFLT - No Free Lunch Theorem; Wolpert and Macready, 1995 şi 1997) arata ca este obligatorie stabilirea unui domeniu de aplicabilitate pentru un euristic
- Categorii de probleme dificile:
  - de decizie, de clasificare (ex. clasificarea filogenetica), de optimizare (ex. a semaforizarii in Cluj), de simulare (ex. a evolutiei speciilor)

Ce sunt algoritmii genetici

- Euristici adaptivi bazați pe ideile teoriei evoluției;
- Elementele cheie la care se face apel:
  - Modelul genetic (dualismul genotip - fenotip) ca in (Morgan, 1915; Fisher, 1918);
  - Încrucişarea (dualismul caractere - gene) ca in (Lamarck, 1830; Mendel, 1865; Weismann, 1893);
  - Mutația: întâmplătoare (De Veies, 1902); deliberată prin expunerea la anumite condiții (Patterson, 1928; Auerbach şi alții, 1947); sub presiunea factorilor de mediu: (Cains şi alții, 1988); etc
- Selecția naturală: supraviețuirea celui mai tare (Darwin, 1859)

Un algoritm genetic



Din “Jäntschi and Bolboacă, A genetic algorithm for structure-activity relationships: software implementation”, manuscript, 26 Jun 2009, http://arxiv.org/abs/0906.4846

Domeniul sau de aplicabilitate

- Relatii structura-activitate:
  - Coreleaza activitatea biologica cu structura chimica
  - Necesita serii de compusi potenti biologic
  - A fost aplicat pe log(K_{ow}) al PCB
- Familii de descriptori de structura:
  - Poseda un cod genetic (genom; gene)
  - Creeaza o populatie ce caracterizeaza structura
  - Creeaza o problema complexa de optimizare
  - A fost aplicat pe familia MDF (Jäntschi, 2004; Jäntschi, 2005; Jäntschi and Bolboacă, 2007)

Evolutie

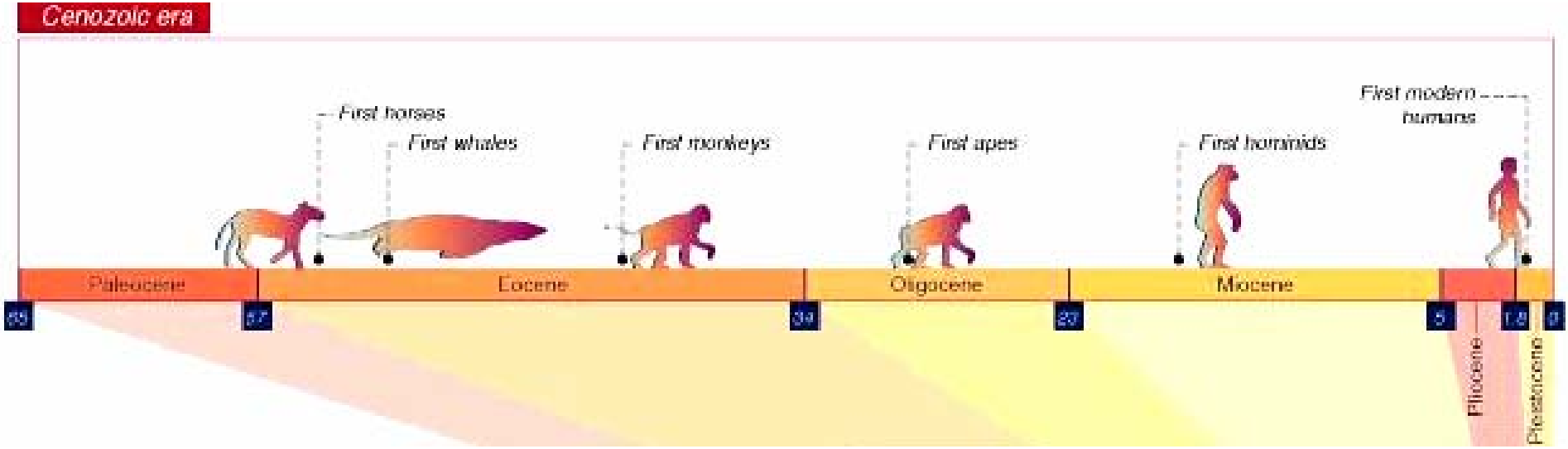
- Optimizare => Simulare
  - Genetic Algorithms in Optimization, Simulation and Modeling (Stender, Hillebrand, and Kingdon, 1994)
- lex parsimoniae: entia non sunt multiplicanda praeter necessitatem (William of OCKHAM, 1288 - 1348)
  - regula lamei de ras a lui Ockham: trebuie eliminate toate acele presupuneri care nu fac nici o diferență în predicțiile observate ale ipotezelor explicatoare sau teoriei
- Optimizare: evolutia MLR folosind un esantion din MDF pe $\log(K_{OW})$ al PCB avand ca obiectiv "r^2=max"
- Simulare: 46 de executii independente ale evolutiei la contingenta (Proportional, Turnir, Deterministic) strategiilor de Selectie vs. Supravietuire

Cadrul evolutiei

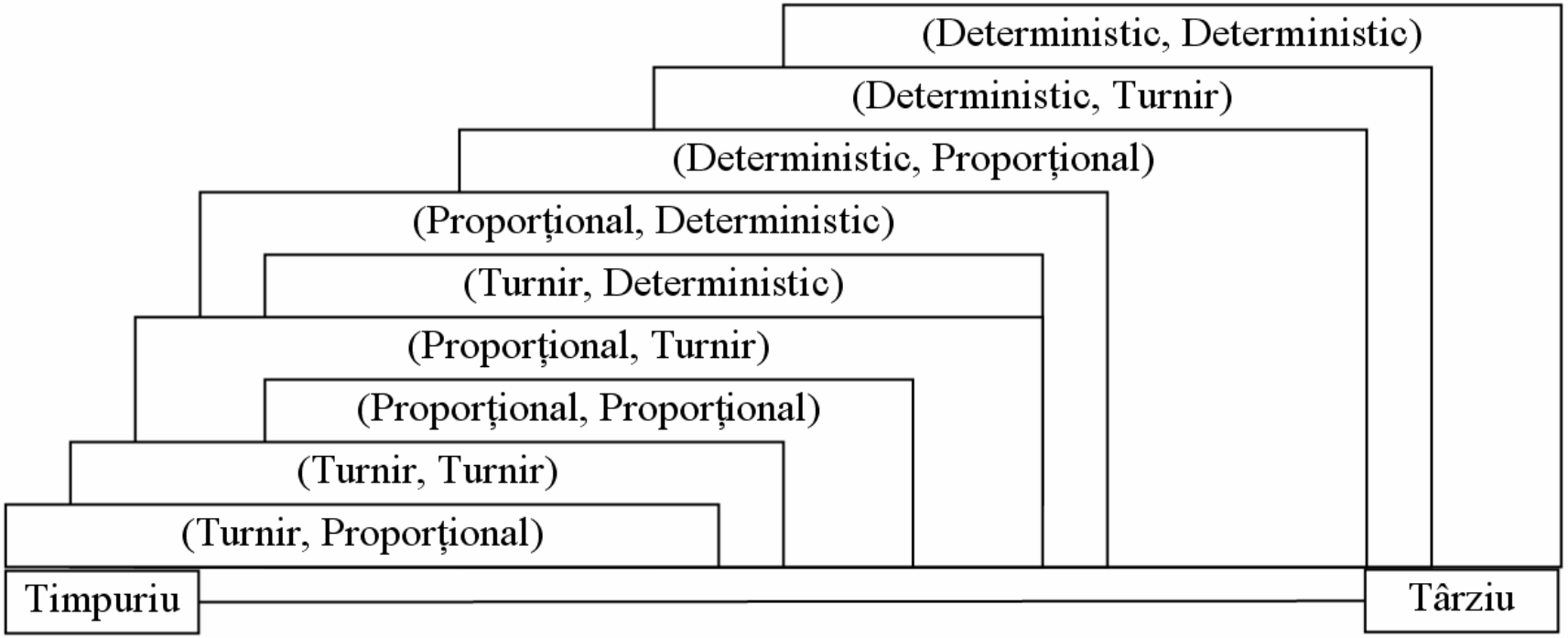
- Obiectiv (fenotipic):
  - r^2(Act.Biol;MLR(4 indivizi MDF))=max.
- Cultivar: max. 6X12=72 indivizi MDF (12 genotipuri)
- Strategia de Evolutie:
  - Cate 4 genotipuri (2 perechi) creeaza descendenti (mutatie si incrucisare) in fiecare generatie
- Strategia de Selectie (pentru Mutatie si Incrucisare):
  - Scor de selectie individ (Obiectiv → max): min{r2(Act.Biol;MLR(individ,3MDF)}
  - Modalitate: Proportional, Turnir, Deterministic
- Strategia de Supravietuire (pentru inlocuirea materialului genetic crescut in Cultivar):
  - Scor de supravietuire individ: inversul mediei de similaritate genotipica si similaritate fenotipica
    - Similaritate genotipica: distante Manhattan intre genotipuri
    - Similaritate fenotipica: distante Manhattan intre scorurile de selectie
  - Modalitate: Proportional, Turnir, Deterministic
- Modalitatati de extragere:
  - Proportional: Sansa de calificare proportionala cu scorul
  - Turnir: Concureaza cate doi alesi la intamplare, se extrage cel mai tare
  - Deterministic: Cel mai tare dintre toti este extras

Care e viteza evolutiei?

- Calculul vitezei de evoluție medie ca raport dintre numărul de evoluții şi numărul de generații arată că ordinul de mărime al vitezei de evoluție (în cazul studiat şi pentru parametrii de experiment definiți) este de 1.37‰ adică în medie 1.37 evoluții la 1000 de generații

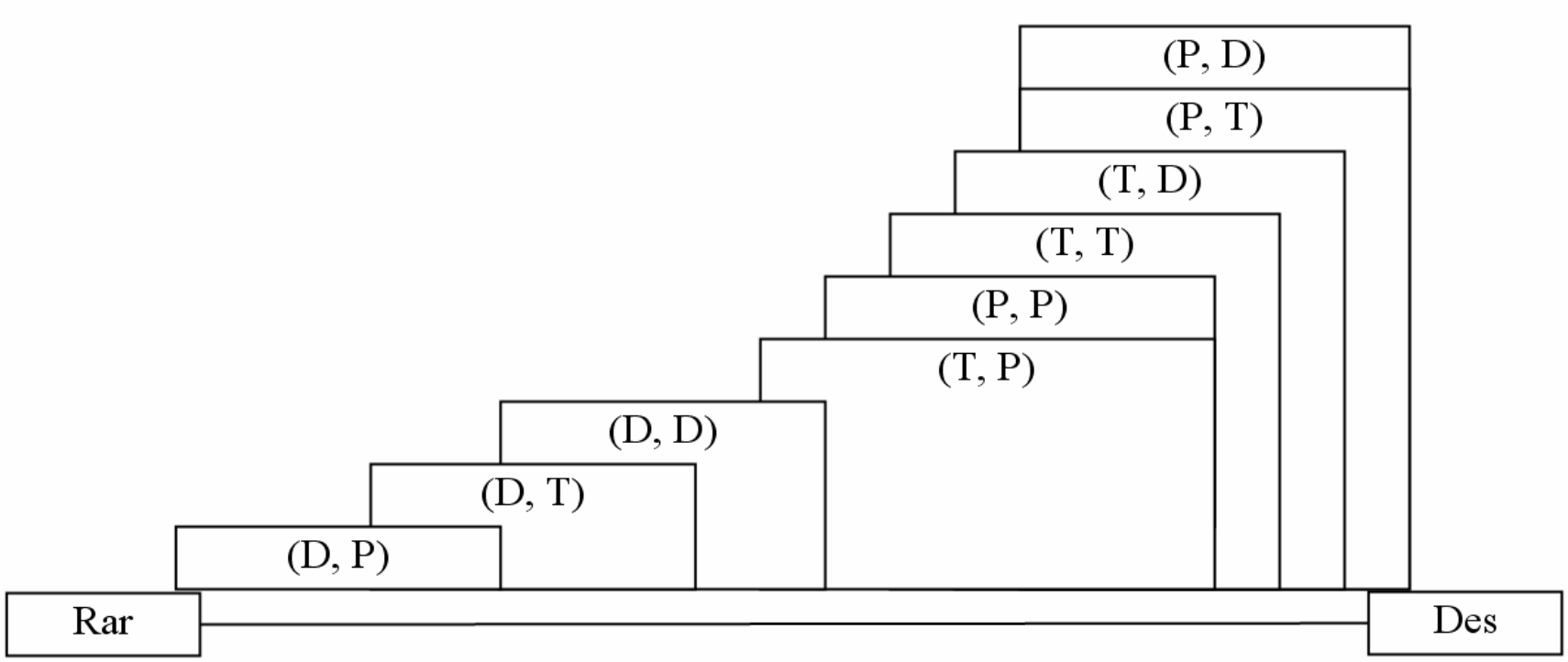


Cât de timpuriu se produc evoluțiile?



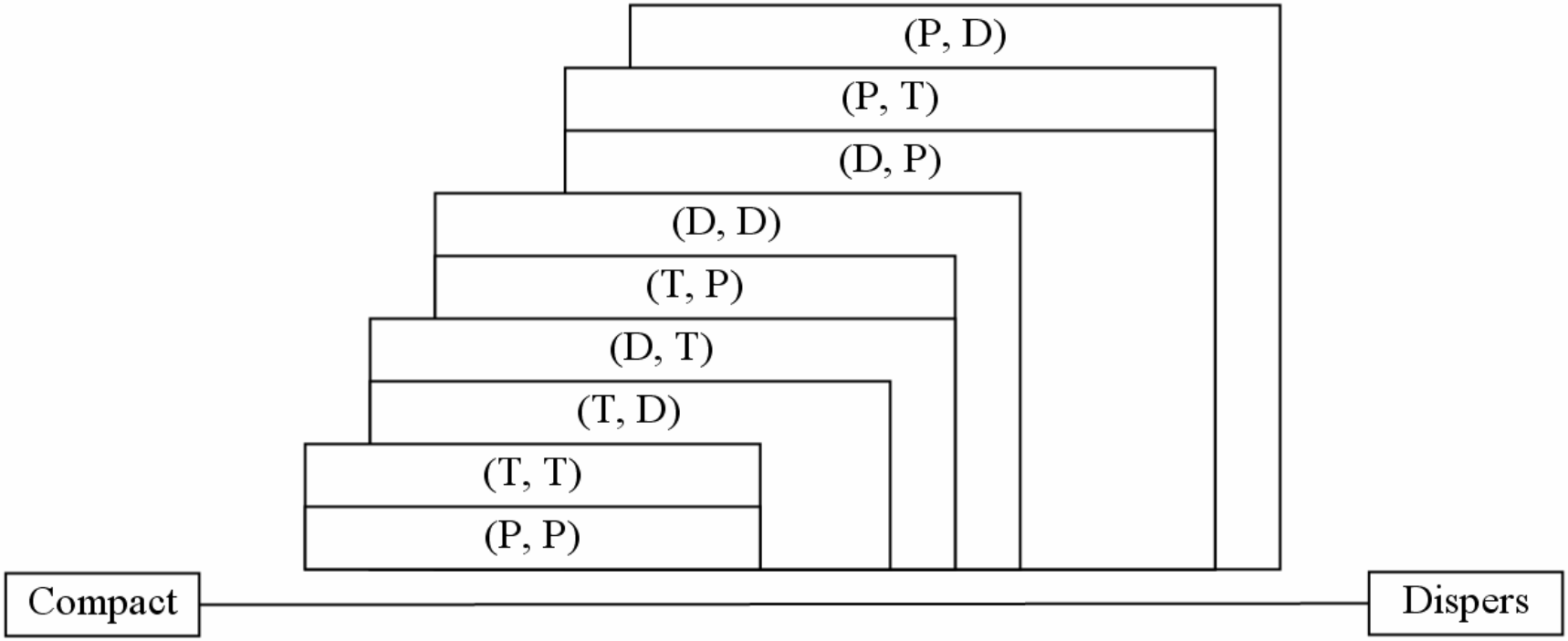
(Selecție, Supraviețuire) [CI(95%,Medie$_{n=46}$(generație medie))]

Cât de frecvent se produc evoluțiile?



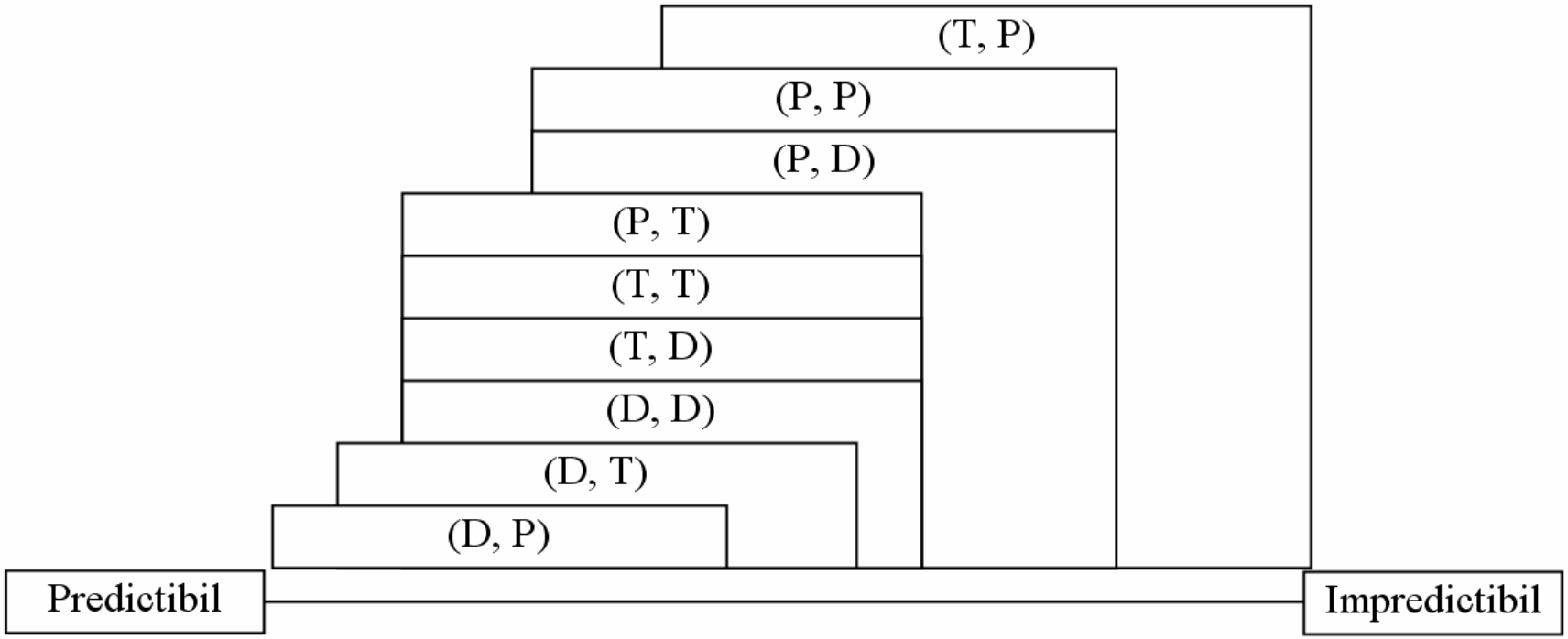
(Selecție, Supraviețuire) [CI(95%,Medie$_{n=46}$(număr evoluții))]

Cât de dispers se produc evoluțiile?



(Selecție, Supraviețuire) [CI(95%,Deviație$_{n=46}$(generație medie))]

Cât de predictibil se produc evoluțiile?



(Selecție, Supraviețuire) [CI(95%,Deviație$_{n=46}$(număr evoluții))]

Distributia evolutiei

- Evolutia este supusa unor factori de stres:
  - Spatiul limitat (max. 12 genotipuri reprezentate de indivizi in cultivar)
  - Crearea de descendenti (prin selectie pentru mutatie si incrucisare) este oferita doar celor mai bune 4 genotipuri
  - Ramanerea in spatiu (prin supravietuire) le este refuzata celor mai slabe 4 genotipuri
- Este de asteptat ca distributia scorurilor obiectiv ale unei generatii (estimata din cele 46 de executii independente) sa fie o distributie de valori extreme (nu distributie Normala – Gauss !!)
- S-a verificat (ca alternativa cu alte 45 legi de distributie) si demonstrat (semnificativ statistic, risc mai mic de 5% de a fi in eroare) ca evolutia (supusa factorilor de stres) creeaza o (sub)populatie distibuita dupa legea Fisher-Tippett (sau generalizata a valorilor extreme GEV) pe parcursul intregii evolutii si imediat dupa primele generatii

Fisher-Tippet (Generalized Extreme Value)

$$FT_{PDF}(X)=\begin{cases}\frac{1}{\beta}\exp\left(-(1+k\frac{x-\lambda}{\beta})^{-1/k}\right)(1+k\frac{x-\lambda}{\beta})^{-1-1/k}, & k<0 \quad \text{Weibull}\\ \frac{1}{\beta}\exp\left(-\frac{x-\lambda}{\beta}-\exp(-\frac{x-\lambda}{\beta})\right), & k=0 \quad \text{Gumbel}\\ \frac{1}{\beta}\exp\left(-(1+k\frac{x-\lambda}{\beta})^{-1/k}\right)(1+k\frac{x-\lambda}{\beta})^{-1-1/k}, & k>0 \quad \text{Fréchet}\end{cases}$$

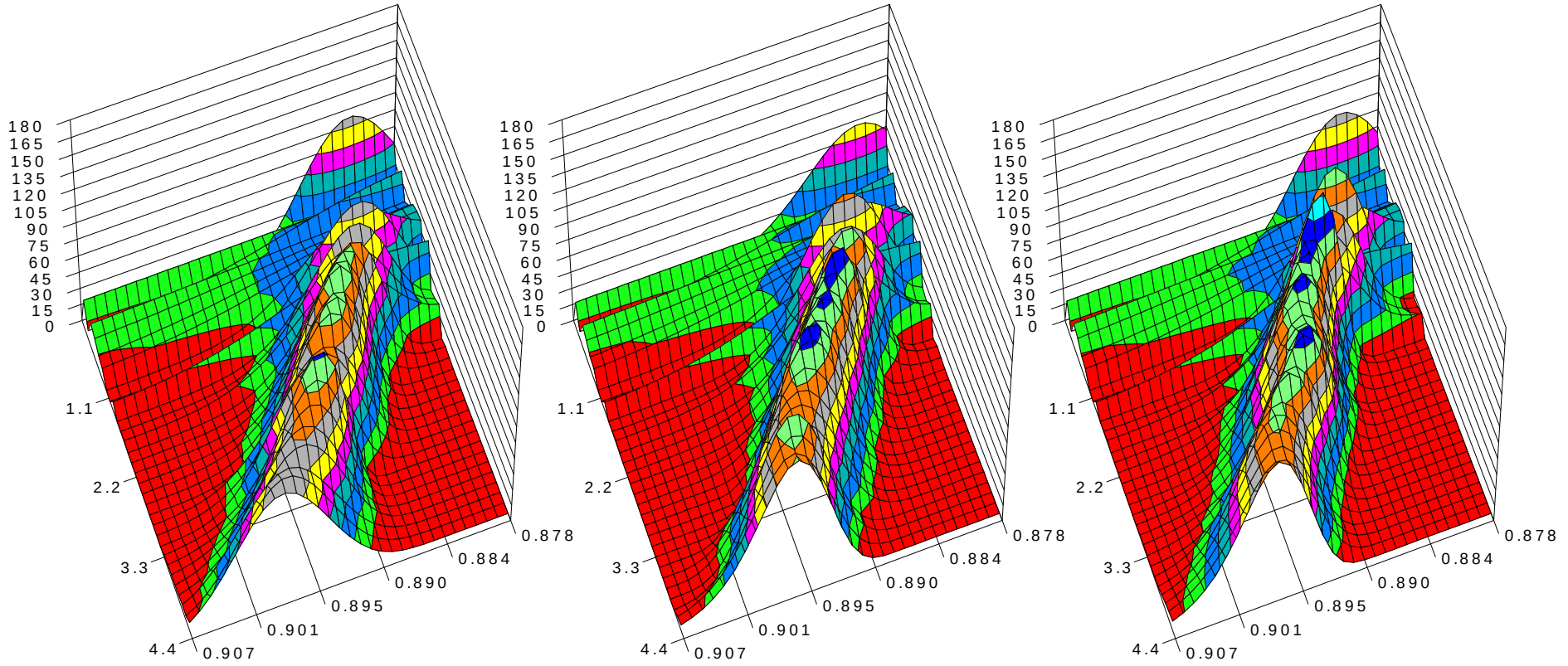
$$FT_{CDF}(X)=\left\{\begin{cases}\exp\left(-(1+k\frac{x-\lambda}{\beta})^{-1/k}\right), & k<0 \quad \text{Weibull}\\ \exp\left(-\exp(-\frac{x-\lambda}{\beta})\right), & k=0 \quad \text{Gumbel}\\ \exp\left(-(1+k\frac{x-\lambda}{\beta})^{-1/k}\right), & k>0 \quad \text{Fréchet}\end{cases}\right\}\text{Fisher}-\text{Tippett}$$

Lege de distributie introdusa in 1928 de Ronald A FISHER si Leonard HC TIPPETT

FISHER RA, TIPPETT LHC. 1928. Limiting Forms of the Frequency Distribution of the Largest of Smallest Member of a Sample. Proceedings of the Cambridge Philosophical Society 24:180-190

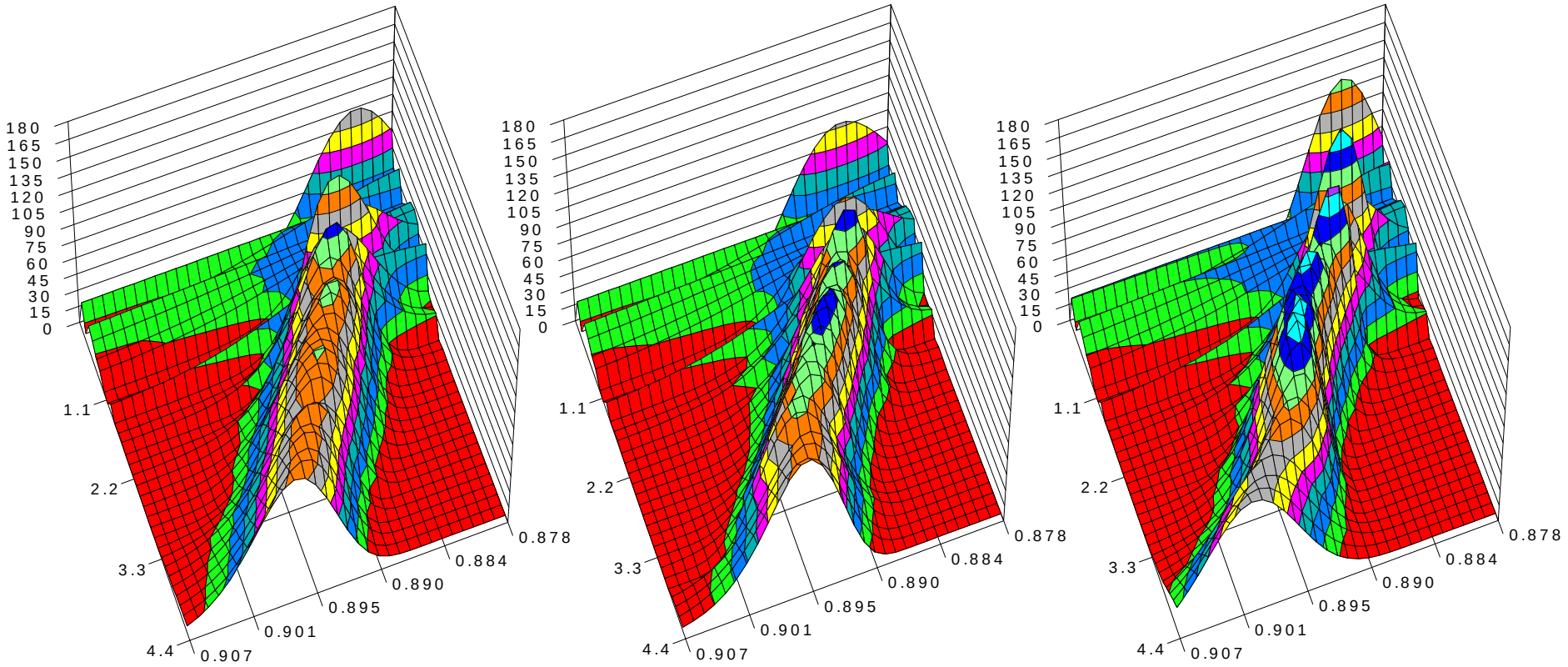
Distributii F-T(GEV) observate - $FT_{PDF}(r^2, \log_{10}Generatie)$

Strategiile cu selectie proportionala: PP, PT si PD



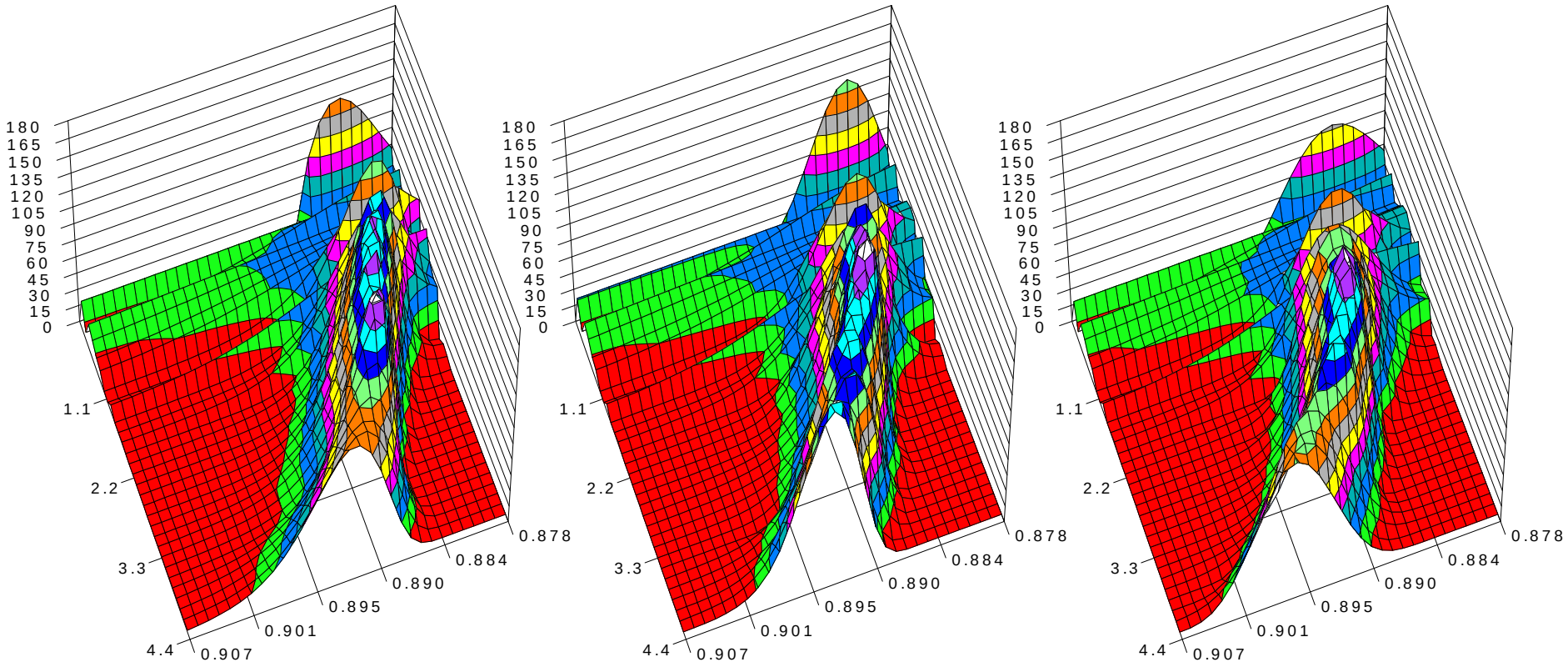
Distributii F-T(GEV) observate - $FT_{PDF}(r^2, \log_{10}\text{Generatie})$

Strategiile cu selectie in turnir: TP, TT si TD

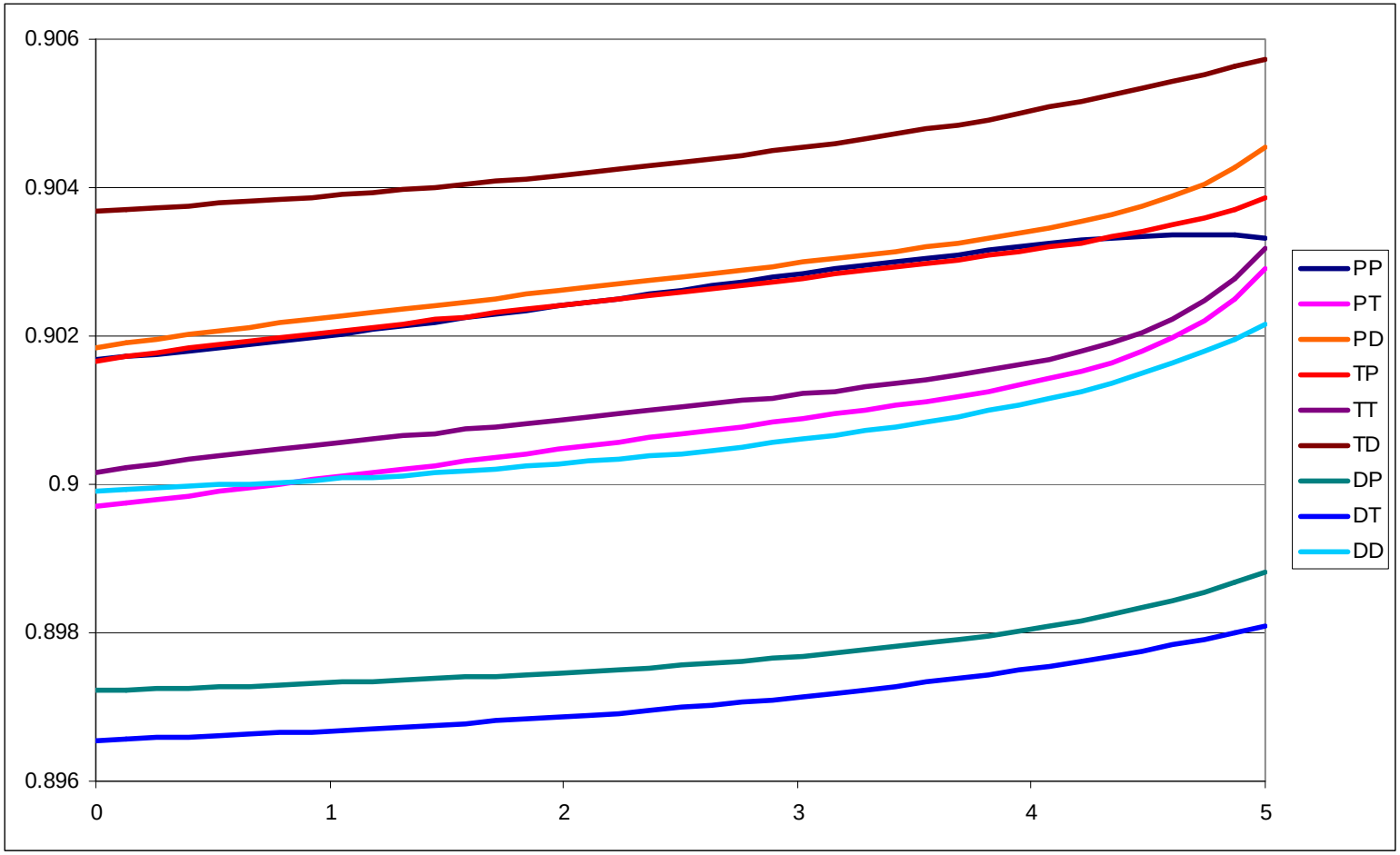


Distributii F-T(GEV) observate - $FT_{PDF}(r^2, \log_{10}\text{Generatie})$

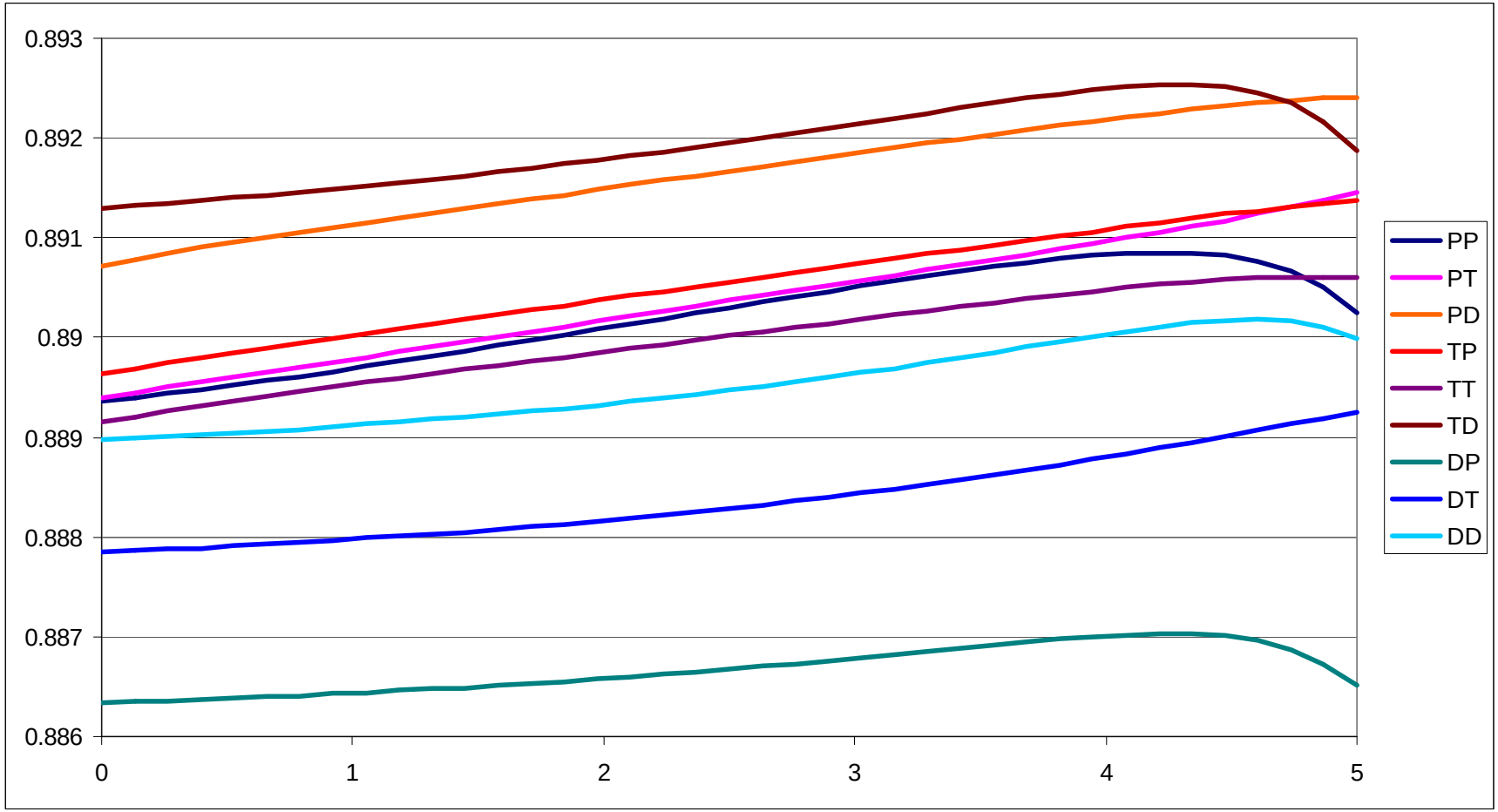
Strategiile cu selectie determinista: DP, DT si DD



Loteria norocosilor (CDF_{FT}=95%)



Loteria norocosilor (CDF_{FT}=5%)



Lărgimea intervalului de normalitate în selecție şi supraviețuire (fără noroc, fără ghinion)