



Общая схема работы генетического алгоритма

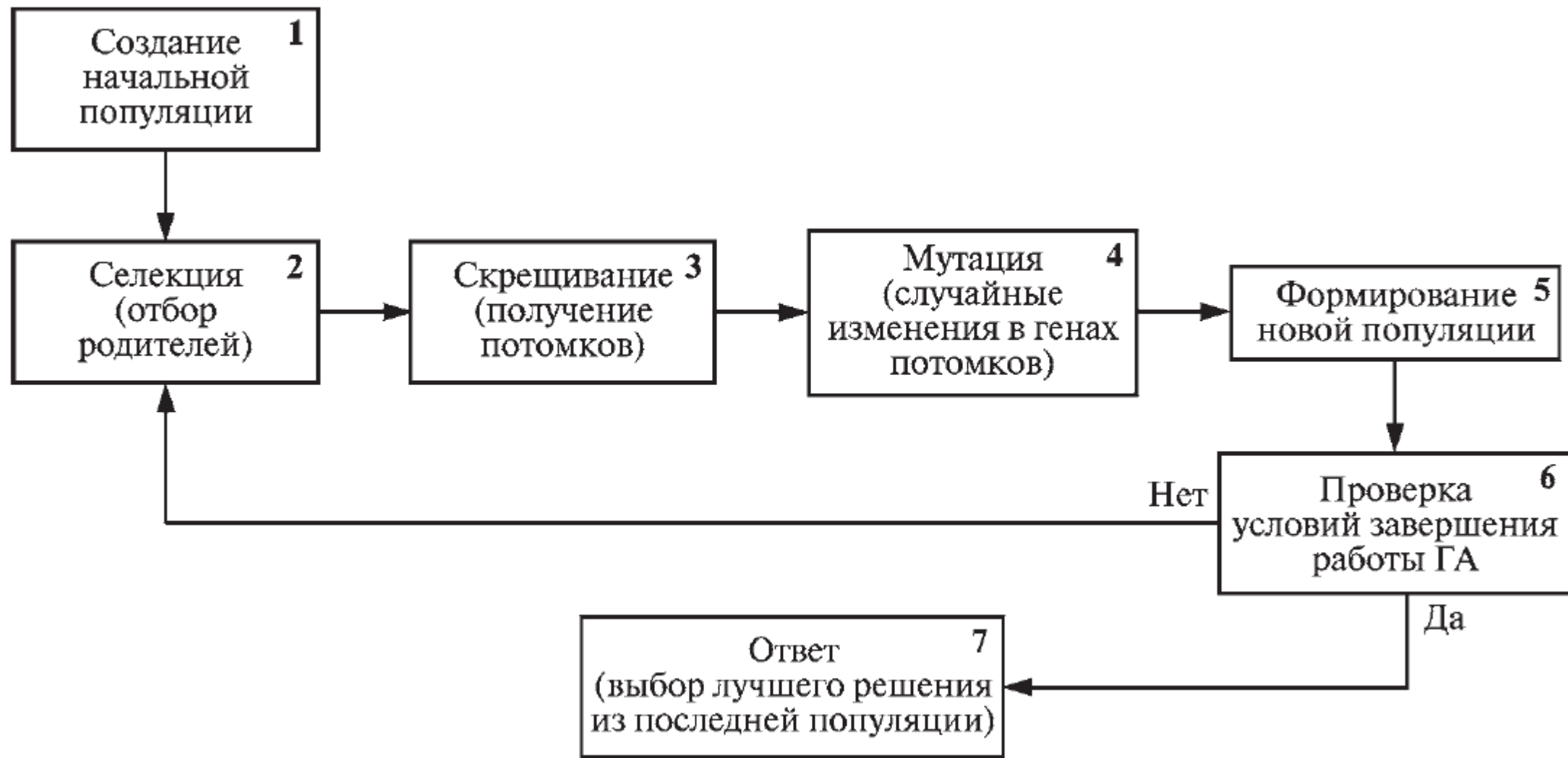
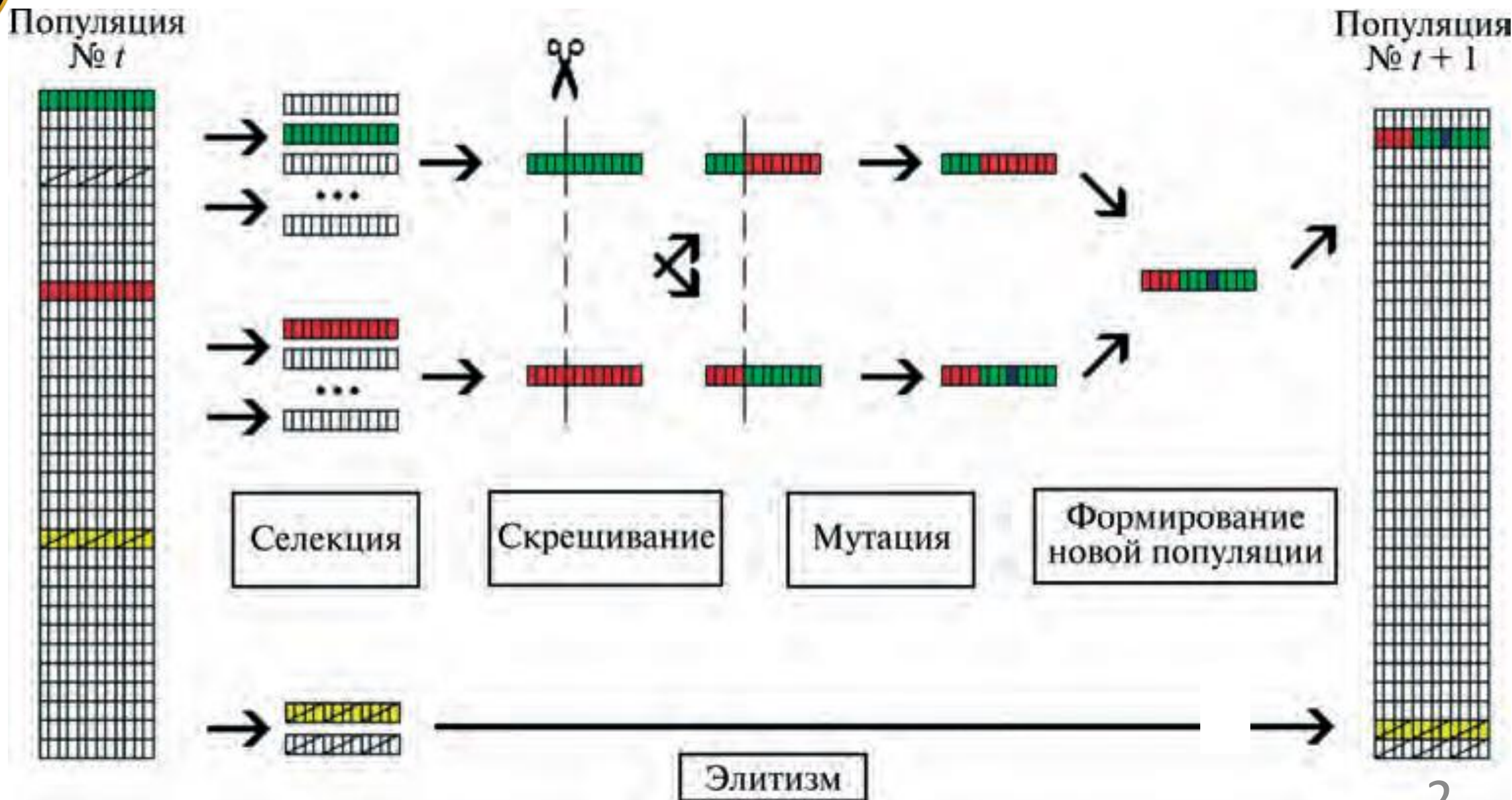




Схема формирования новой популяции



Настроечные параметры генетических алгоритмов:

- число поколений,
- количество особей в популяции,
- способ отбора родителей (вид селекции), число участников турнира,
- вероятность мутации,
- стратегия отбора в новую популяцию.

Способ отбора родителей (вид селекции)

- случайный равновероятный отбор (все особи имеют равные шансы стать родителями);
- рангово-пропорциональный отбор (родителями могут стать особи, входящие в $k\%$ лучших по значениям функции приспособленности);
- рулетка (вероятность особи быть отобранной в родители пропорциональна значению ее функции приспособленности);
- турнирный отбор (случайным образом выбирается определенное число особей, среди которых побеждает особь с максимальным значением функции приспособленности).



Символьная регрессия. Идея метода

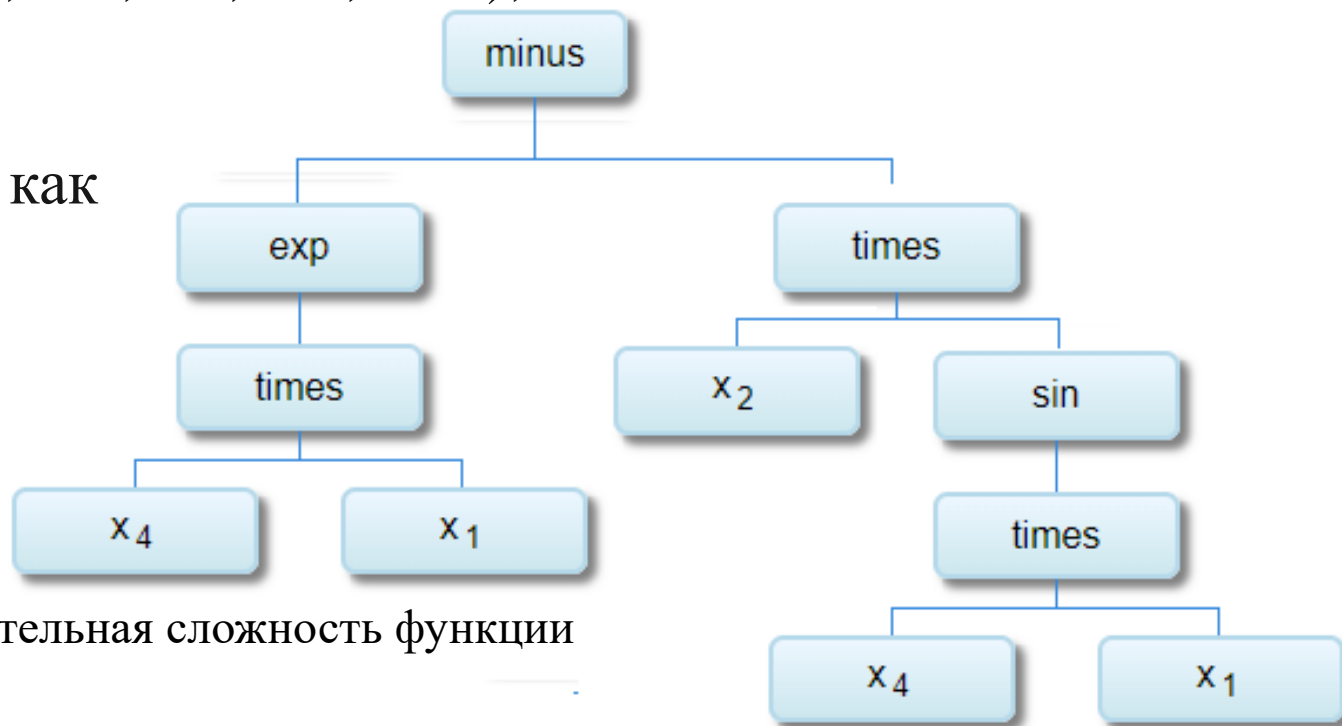
Гены:

- переменные,
- арифметические операции, включая унарный минус,
- функции (log, exp, sqrt, ^2, ^3, abs, sin, cos, tanh),
- Константы

Любая функция может быть представлена как хромосома, состоящая из набора генов

$$y = \exp(x_1 * x_4) - x_2 * \sin(x_1 * x_4)$$

Листья дерева - входные переменные и константы, узлы – арифметические операции и функции. Вычислительная сложность функции определяется числом узлов и глубиной дерева.

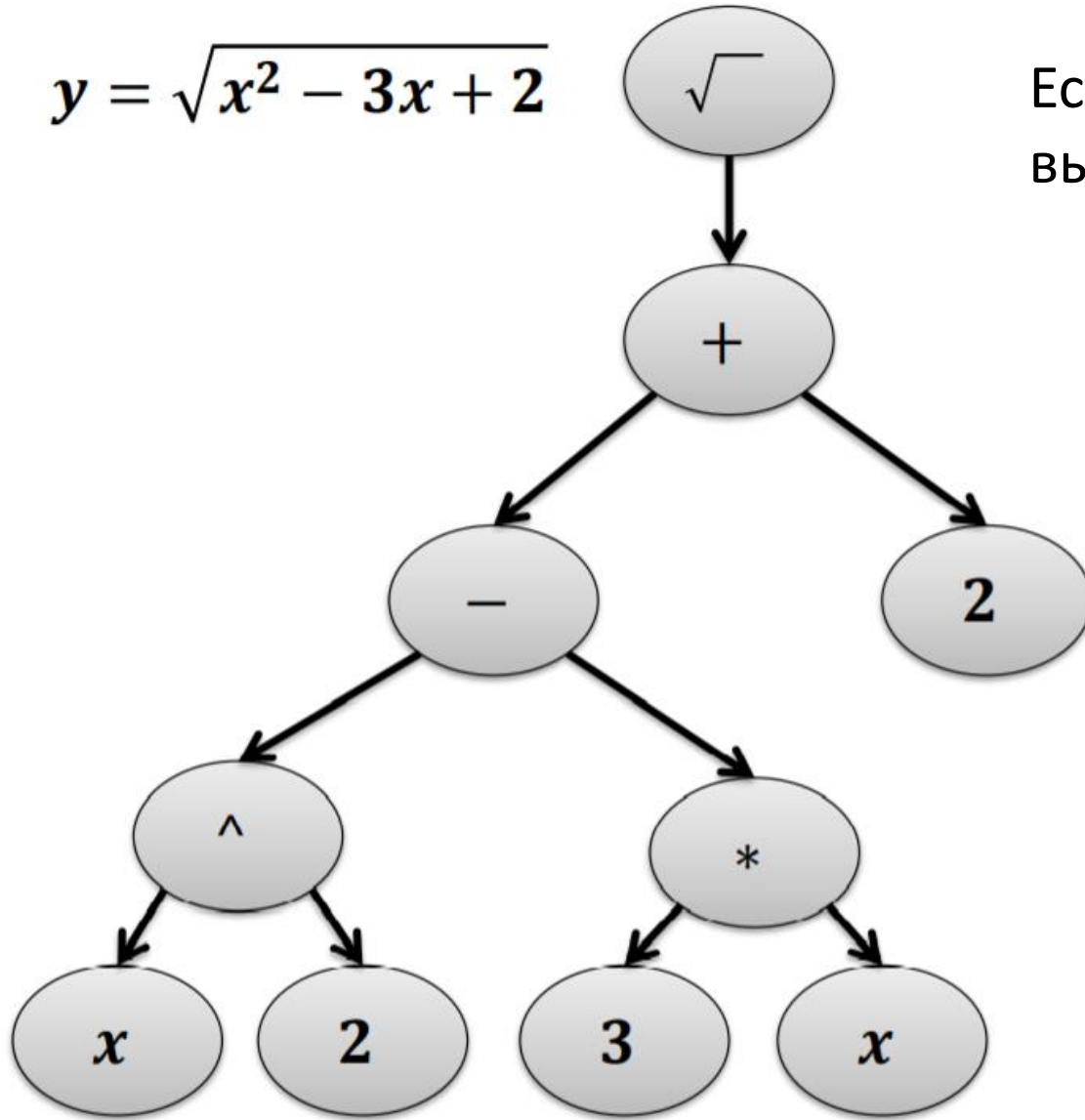


постфиксная запись

x_1	x_4	*	exp	x_2	x_4	x_1	*	sin	*	-
-------	-------	---	-----	-------	-------	-------	---	-----	---	---

Синтаксическое дерево

$$y = \sqrt{x^2 - 3x + 2}$$



Есть ли однозначное соответствие между выражением и синтаксическим деревом?

Обратная польская запись (постфиксная)

Reverse Polish notation

Все аргументы (операнды) расположены перед знаком операции.

Порядок выполнения операций однозначно задаётся порядком следования знаков операций в выражении, поэтому отпадает необходимость использования скобок и введения приоритетов и ассоциативности операций.

Запись набора операций состоит из последовательности операндов и знаков операций.

Операнды в выражении при письменной записи разделяются пробелами.

Выражение читается слева направо. Когда в выражении встречается знак операции, выполняется соответствующая операция над двумя последними встретившимися перед ним операндами в порядке их записи. Результат операции заменяет в выражении последовательность её операндов и её знак, после чего выражение вычисляется дальше по тому же правилу.

Результатом вычисления выражения становится результат последней вычисленной операции

Обратная польская запись

Reverse Polish notation

Традиционная (инфиксная) запись

$$2 * (3 + 5) - (6 + 7) / (8 - 9)$$

Вычисление выражения

$$3 + 5 = 8; 8 * 2 = 16$$

$$6 + 7 = 13; 8 - 9 = -1$$

$$13 / -1 = -13;$$

$$16 - (-13) = 29$$

Обратная польская (постфиксная) запись

3 5 + 2 * 6 7 + 8 9 - / -

Вычисление выражения

Вход	3 5 + 2 * 6 7 + 8 9 - / -
Стек	3 5
Вход	+ 2 * 6 7 + 8 9 - / -
Стек	8 2
Вход	* 6 7 + 8 9 - / -
Стек	16 6 7
Вход	+ 8 9 - / -
Стек	16 13 8 9
Вход	- / -
Стек	16 13 -1
Вход	/ -
Стек	16 -13
Вход	-
Стек	29

Обратная польская запись

1. Обработка входного символа

Если на вход подан операнд, он помещается на вершину стека.

Если на вход подан знак операции, то операция выполняется над требуемым количеством значений, извлечённых из стека, взятых в порядке добавления. Результат выполненной операции кладётся на вершину стека.

2. Если входной набор символов обработан не полностью, перейти к шагу 1. Иначе - результат вычисления выражения лежит на вершине стека.

Обратная польская (постфиксная) запись



Вход	3 5 + 2 * 6 7 + 8 9 - / -
Стек	3 5
Вход	+ 2 * 6 7 + 8 9 - / -
Стек	8 2
Вход	* 6 7 + 8 9 - / -
Стек	16 6 7
Вход	+ 8 9 - / -
Стек	16 13 8 9
Вход	- / -
Стек	16 13 -1
Вход	/ -
Стек	16 -13
Вход	-
Стек	29

Обратная польская запись

Базовые операции

1. Поместить в стек x
2. Поместить в стек число
3. Сложение и вычитание
4. Умножение и деление (в виде $x/(y + \delta)$)
5. Возведение в степень (в виде $|x|^y$)
6. Унарный минус

Внимание! Они должны возвращать корректный результат при любых входных данных!

Обратная польская (постфиксная) запись

3 5 + 2 * 6 7 + 8 9 - / -

Вычисление выражения

Вход	3 5 + 2 * 6 7 + 8 9 - / -
Стек	3 5
Вход	+ 2 * 6 7 + 8 9 - / -
Стек	8 2
Вход	* 6 7 + 8 9 - / -
Стек	16 6 7
Вход	+ 8 9 - / -
Стек	16 13 8 9
Вход	- / -
Стек	16 13 -1
Вход	/ -
Стек	16 -13
Вход	-
Стек	29

Обратная польская запись

Записать в виде синтаксического дерева и обратной польской записи выражение

$$\frac{\cos^2 x + \sin^2 y}{10^z}$$

Найти значение выражения, записанного в обратной польской записи, и перевести его в обычную (инфиксную) форму записи

$$14\ 2\ 3\ 4\ -\ *\ / 6\ -$$

Символьная регрессия

Члены популяции

Выражения в обратной польской записи

Функция приспособленности

$$RSS = \sum_i (y_i - \hat{y}_i)^2$$

Базовые операции

1. Поместить в стек x
2. Поместить в стек число
3. Сложение и вычитание
4. Умножение и деление (в виде $x/(y + \delta)$)
5. Возведение в степень (в виде $|x|^y$)
6. Унарный минус

Внимание! Они должны возвращать корректный результат при любых входных данных!

Скращивание

Шаг 1. Взять два случайных члена популяции

Шаг 2. Разделить каждое выражение на две части и поменять их местами

$A \ B \ + \ C \ D \ + \ * \Rightarrow (A+B) * (C+D)$

$C \ D \ / \ B \ A \ ^ \ + \Rightarrow (C/D) + B^A$

Результат

$A \ B \ + \ C \ / \ B \ A \ ^ \ + \Rightarrow (A+B) / C + B^A$

$C \ D \ D \ + \ * \Rightarrow C * (D+D)$

Шаг 3. Внести случайные изменения («мутации» в коэффициенты и операции)



Символьная регрессия. Фитнес функция

$$ff_j = - \sum_{i=1}^n (F_j(x_{1i}, x_{2i}, \dots, x_{ni}) - y_i)^2$$

где

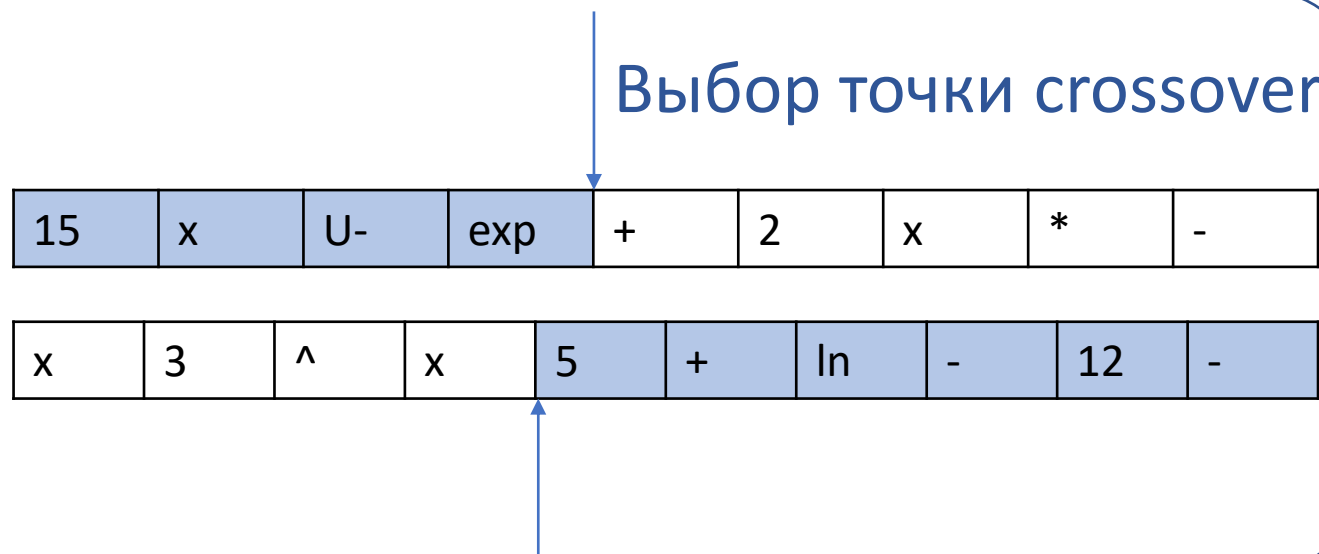
$x_{1i}, x_{2i}, \dots, x_{ni}$ — значения объясняющих переменных для i -ой записи обучающей выборки,
 y_{ni} — известное значение целевой переменной для i -ой записи обучающей выборки,
 n — объем обучающей выборки,
 j — номер особи (сгенерированной функции) .

Перед знаком суммы знак “-“, так как в соответствии с эволюционным принципом выживают наиболее приспособленные особи, и значения функции приспособленности должны стремиться к максимуму.

Символьная регрессия. Пример операций скрещивания и мутации

Родители

a) $15 + \exp(-x) - 2 * x$
b) $x^3 - \ln(x + 5) - 12$



Потомки

15	x	U-	exp	5	+	ln	-	12	-
----	---	----	-----	---	---	----	---	----	---

x	3	^	x	+	2	x	*	-
							^	

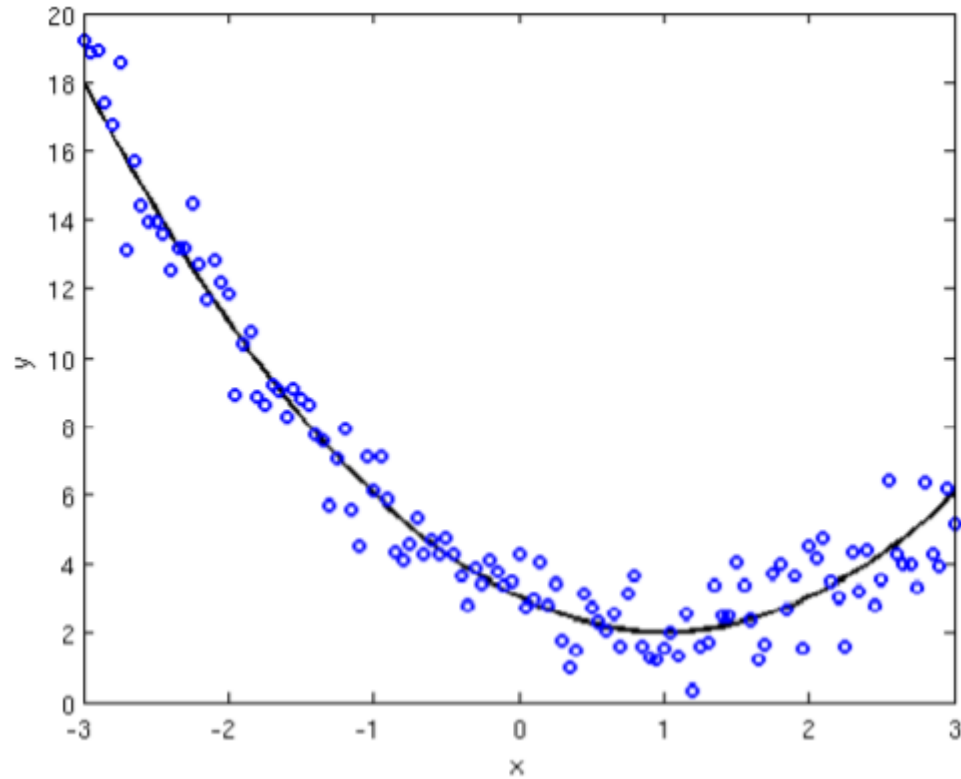
Мутация гена

a') $15 - \ln(\exp(-x) + 5) - 12$

b') $x^3 + x - 2x$

b'_m) $x^3 + x - 2^x$

Пример символьной регрессии



$$y = x^2 - 2x + 3$$

Результат регрессии в обратной польской записи

`x [-1.0409] + u- [2.0115] ^ [3.8315] [0.3417] ^ [0.6995] / [0.6995] ^ +`

Формула после преобразований

$$\begin{aligned} & (x - 1.0409)^{2.0115} + (3.8315^{0.3417} / 0.6995)^{0.6995} = \\ & = (x - 1.0409)^{2.0115} + 1.7701 \end{aligned}$$

Пример результата символьной регрессии для расчета коэффициента сжимаемости

Формула	R^2	R^2_{adj}	WAPE, %
$-2.23887 \cdot 0.00283^{0.05491x_3} x_1^{0.05491x_3} + 0.20215x_5^{9.02 \cdot 10^{-13} \left(\frac{x_1}{x_2 x_3} \right)} - 19.88388$ $+ \frac{23.15947}{\left(\ln \left(\frac{x_2}{14.00667 \ln(x_2)} \right) + 1.89112 \right)^{\frac{x_3}{x_1 - 12.47056}}} - \frac{32.25716}{x_3 - 35.62864} - \frac{460.77293}{x_2}$	0.9638	0.9637	1.25
$329.944x_4 - 0.074x_5^{\left(\frac{\ln(x_2)}{x_1 + \ln(x_1) + 28.104} \right)} - 115.488(x_3 + x_4)^{x_4} + 112.088$ $+ 1.065 e^{\left(-\frac{0.005x_2}{x_1} - x_4 \right)} + \frac{3.859 \left(\frac{x_4}{x_1} \right)}{\left(\frac{18.196}{x_1} \right) + \frac{361,653}{x_1 e^{x_2}}} + \frac{0.031 \ln(x_2)}{10.316x_1 x_5 \ln(x_5) + 18.945} - \frac{59.306}{x_2}$	0.9442	0.9439	1.43

GPTIPS 2MATLAB: an open-source software platform for symbolic data mining Dominic P. Searson

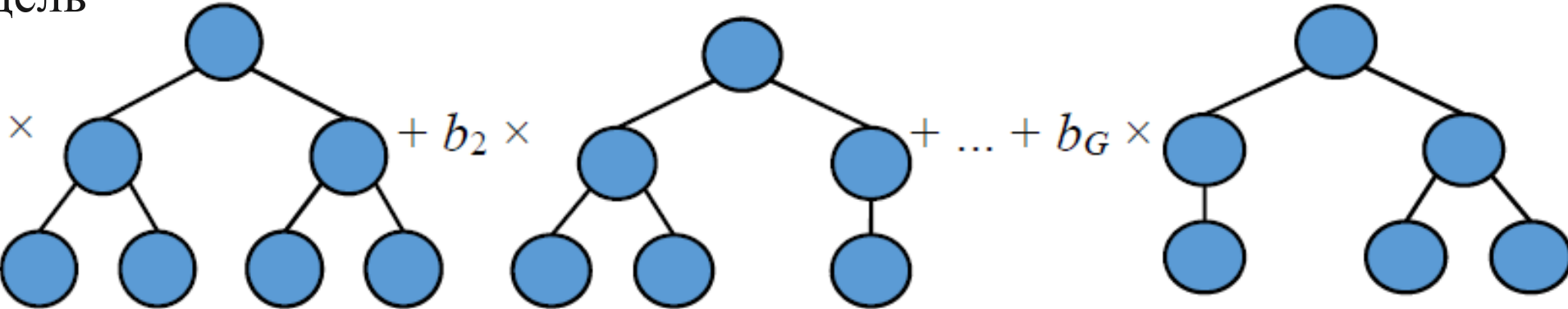
Свободно распространяемое расширение для Матлаб

Компактные формулы, включающие линейные и нелинейные слагаемые,
контроль за сложностью модели



Модификация ГА – мультигенное генетическое программирование

Рассматривается линейная комбинация деревьев-генов с весами
множественная регрессия, где входные переменные – структура типа дерева,
которые являются нелинейными функциями входных переменных, псевдо-
линейная модель

$$\hat{y} = b_0 + b_1 \times \text{tree}_1 + b_2 \times \text{tree}_2 + \dots + b_G \times \text{tree}_G$$


$$\hat{y} = b_0 + b_1 \mathbf{t}_1 + \dots + b_G \mathbf{t}_G$$

\mathbf{t}_i – вектор $(N \times 1)$ значений i -го гена – дерева,

\mathbf{G} – матрица $(N \times (G + 1))$ $\mathbf{G} = [\mathbf{1} \ \mathbf{t}_1 \ \dots \ \mathbf{t}_G]$

$$\hat{\mathbf{y}} = \mathbf{G}\mathbf{b}$$

Оценки коэффициентов b_0, b_1, \dots, b_G (вектор $((G + 1) \times 1)$) могут быть получены с помощью МНК из обучающей выборки: $\mathbf{b} = (\mathbf{G}^T \mathbf{G})^{-1} \mathbf{G}^T \mathbf{y}$

(используется псевдообращение Мура — Пенроуза)



ГОСТ 30319.3— 2015. Газ природный. Методы расчета физических свойств. Вычисление физических свойств на основе данных о компонентном составе

Рекомендована итеративная процедура решения нелинейного уравнения. Погрешность метода расчета коэффициента сжимаемости 0,1 – 0,2%. Если для аппроксимации построить простую расчетную зависимость с погрешностью в 3-5 раз меньше методической, то можно получить значительный выигрыш при решении задач планирования, идентификации параметров, моделирования нестационарных режимов.



Компонентный состав природного газа

Component	MIN, %	MAX, %
Methane	92.2	98.87
Nitrogen	0.5	0.9
Carbon dioxide	0.025	0.435
Ethane	0.33	4.7
Propane	0.0051	0.95
Isobutane	0.0066	0.26
n-Butane	0.0025	0.24
Isopentane	0.0025	0.045
n-Pentane	0.0025	0.03
n-Hexane	0.0025	0.0124
Hydrogen	0.001	0.0023
Oxygen	0.0046	0.0085
Helium	0.0086	0.0147

Таблица создана на основе фактических данных о составах газа транспортируемых по ЕСГ



Генерация набора данных

2 диапазона давления,
температура 273-333К

нормализация параметров
в диапазоне [0, 1]

Диапазон	Давление, МПа	Объем выборки
P1	3.5-5.6	15 900
P2	5.5-7.5	13 333

$$X' = \frac{X - X^{\min}}{X^{\max} - X^{\min}}$$

Во входные параметры
модели добавлена
молярная масса газа

$$M_m = \sum_{i=1}^{N_c} x_i M_i$$

x_i — молярная доля i -го компонента природного газа,

M_i — молярная масса i -го компонента, N_c — количество компонентов



Примеры сгенерированных формул

R^2	Model complexity	Model
0.999	118	$0.0262 T - 0.0168 P - 0.00804 Mm - 0.00122 T^2 (Mm + P) - 0.00115 (2 Mm + 2 T) (Mm + P + 2 T) + 0.00872 T (Mm + P) + 0.00114 Mm(CO_2 + 2 \text{ Methane} + N + T) + 0.923$
0.999	92	$0.0263 T - 0.0169 P - 0.00844 Mm - 0.00134 T^2 (Mm + P) - 0.00105 (2 Mm + 2 T) (Mm + P + 2 T) + 0.00843 T (Mm + P) + 0.923$
0.998	36	$0.0263 T - 0.0166 P - 0.01 Mm + 0.00639 P T + 0.00214 T (Mm - P T) - 0.00488 T^2 + 0.924$
0.983	6	$0.0265 T - 0.0176 P - 0.00998 Mm - 0.00478 T^2 + 0.924$

Коэффициент детерминации

Коэффициент детерминации – доля разброса данных, объясненная регрессионной моделью.

$$R^2 = 1 - \frac{SS_{остат}}{SS_{общ}} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

$$SS_{общ} = \sum_{i=1}^n (y_i - \bar{y})^2$$

Общий разброс (total sum of squares TSS) значений переменной y_i относительно среднего значения

$$SS_{остат} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Остаточный (необъясненный) разброс (residual sum of squares RSS) – отклонения y_i от модельных значений \hat{y}_i

$$SS_{фактор} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

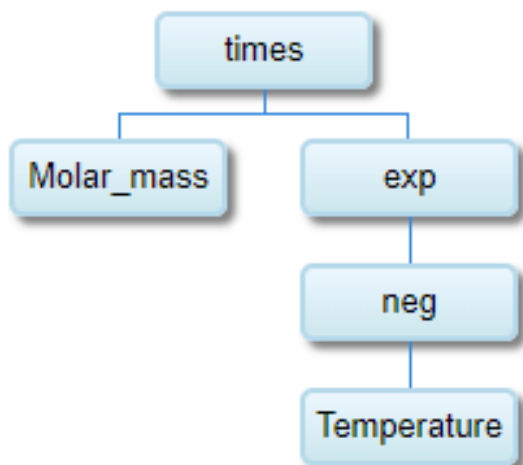
Объясненный (или факторный) разброс (explained sum of squares ESS) – отклонения модельных значений от среднего



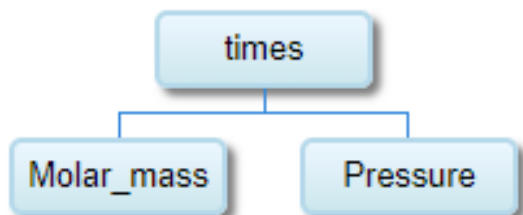
Структура дерева генов модель ID 110

$$Z_{110} = 8.92e-4 \text{ Carbon_dioxide} - 0.00178 \text{ Ethane} + 8.92e-4 \text{ Nitrogen} - 0.0428 P - 8.92e-4 \text{ Propane} - 0.0521 \exp(-T) + 0.0257 P T - 0.0212 Mm \exp(-T) - 0.00934 Mm P + 0.979$$

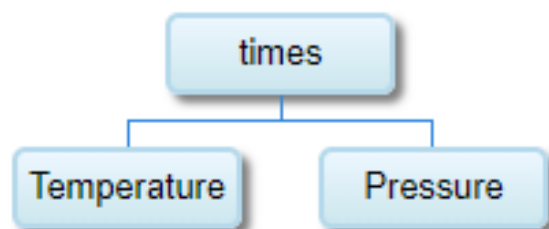
Gene 1



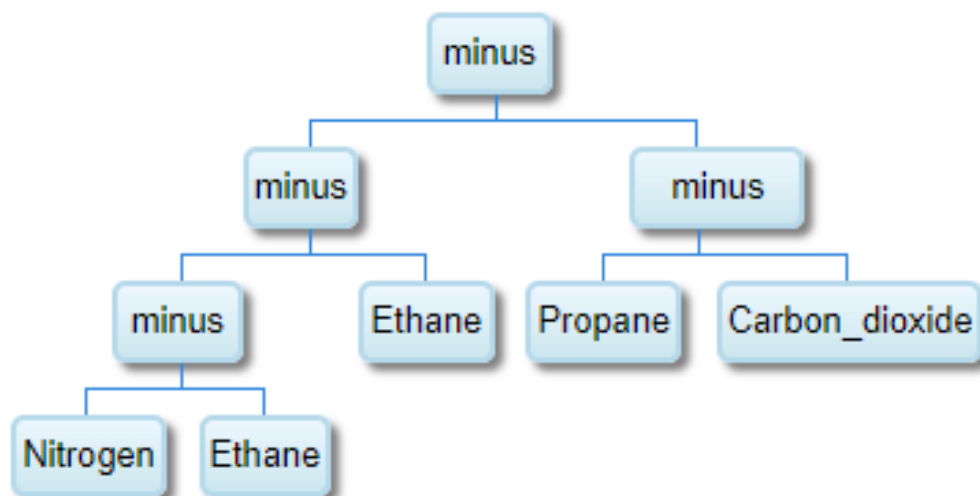
Gene 2



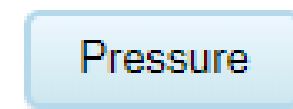
Gene 3



Gene 4



Gene 5



Gene 6





Структура дерева генов модель ID 605

$$Z_{605} = 0.0139 T - 0.00166 \text{ Ethane} - 0.0302 \exp(-T) - 0.019 \text{ Mm} \exp(-T) - 0.0431 P \exp(-T) - 0.0129 \text{ Mm} P + 0.957$$

Gene 1

Temperature

Gene 2

negexp

Temperature

Gene 3

times

Pressure

Molar_mass

Gene 4

times

negexp

Molar_mass

Temperature

Gene 5

times

negexp

Pressure

Temperature

Gene 6

Ethane



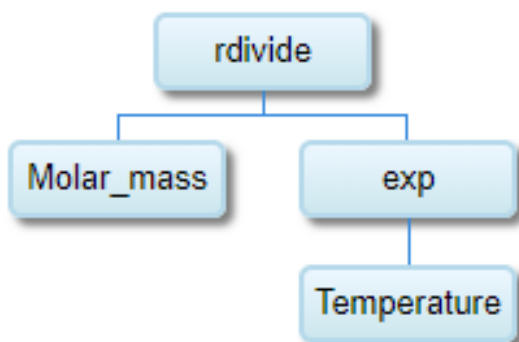
Структура дерева генов модель ID 8

$$Z_g = 0.0016 \text{ Nitrogen} - 0.00283 \text{ Ethane} + 0.0224 T - 0.0381 \exp(-T) - 0.00771 \exp(-2T) - 0.0328 Mm \exp(-T) - 0.0381 P \exp(-T) - 0.0112 Mm P + 0.933$$

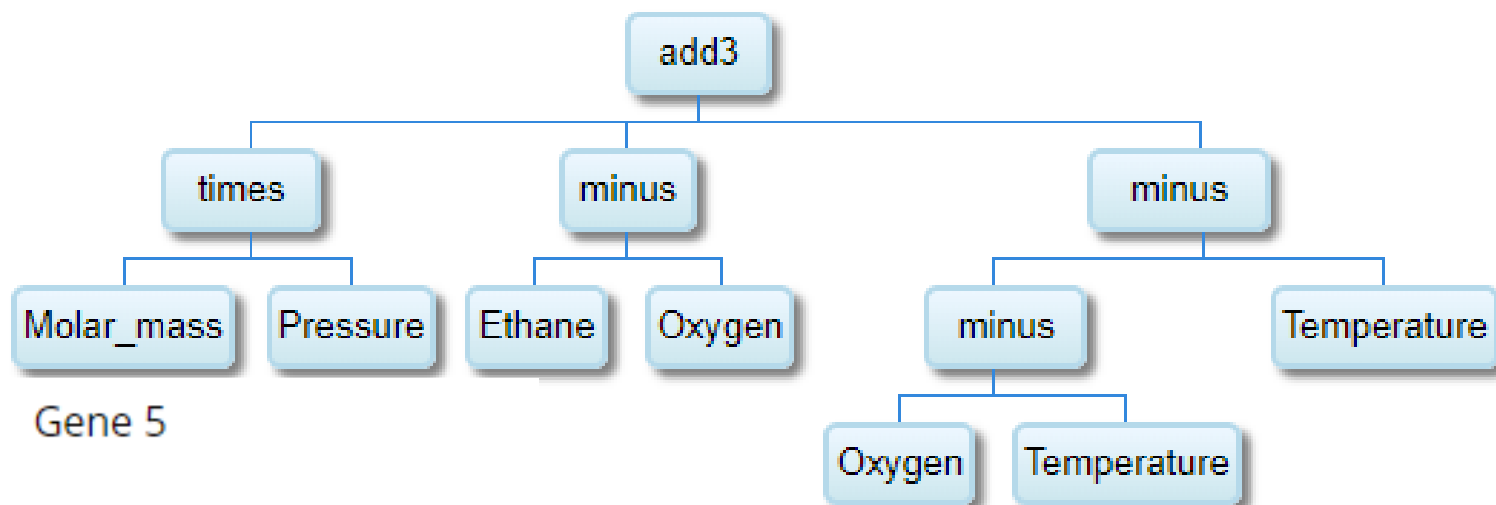
Gene 1

Nitrogen

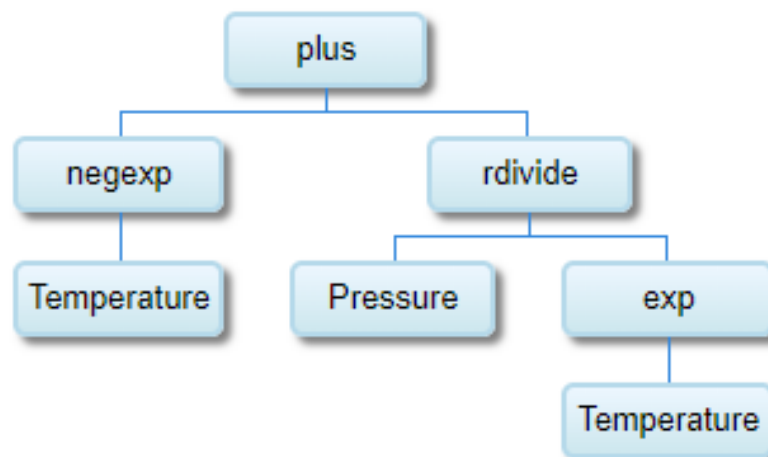
Gene 2



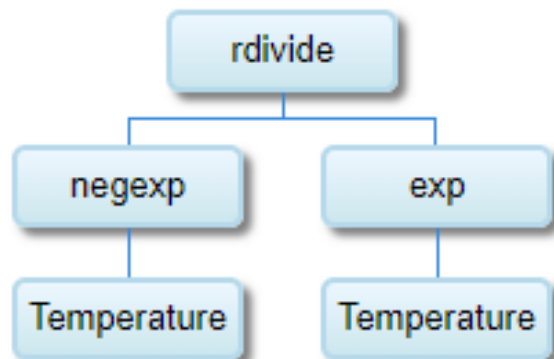
Gene 3



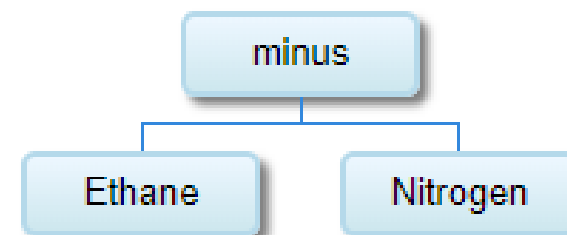
Gene 5



Gene 4



Gene 6



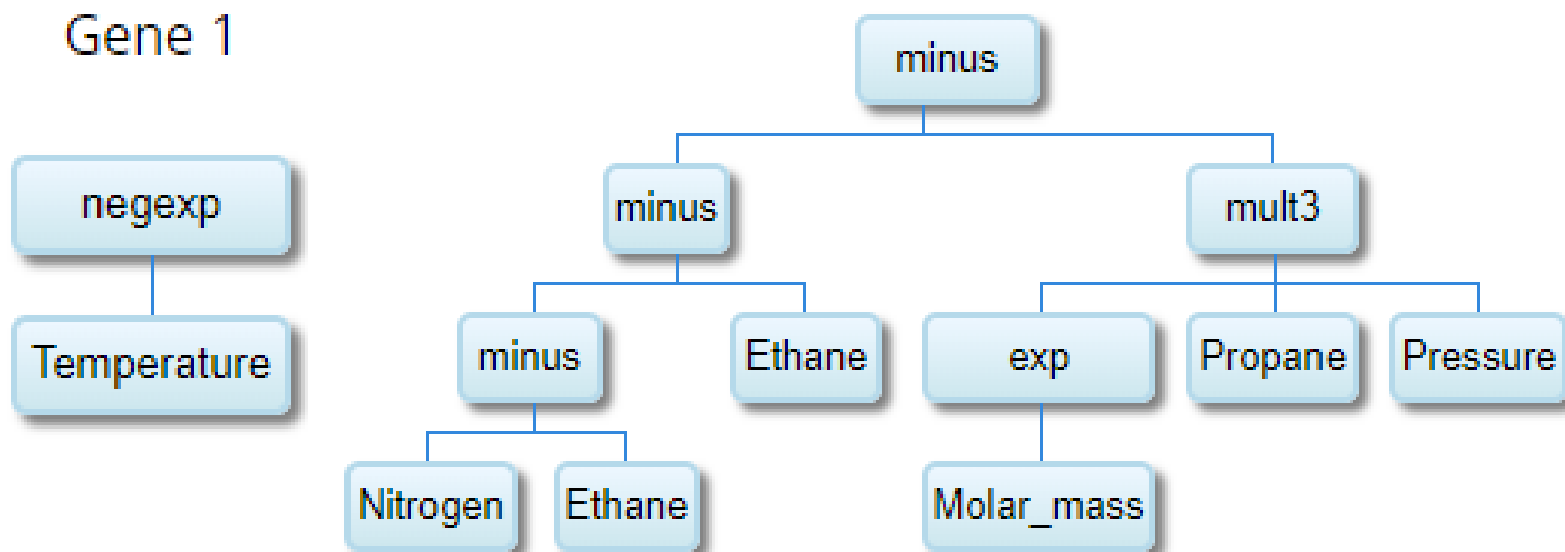


Структура дерева генов модель ID 15

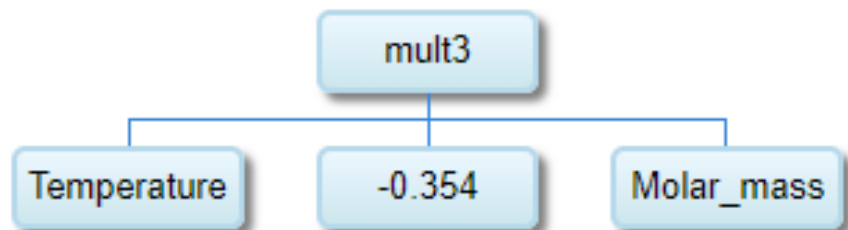
$$Z_{15} = 0.00136 \text{ Nitrogen} - 0.0369 \text{ Mm} - 0.00273 \text{ Ethane} - 0.00559 \text{ P} - 0.0824 \exp(-T) + 0.0226 \text{ Mm} T - 0.0378 \text{ P} \exp(-T) - 0.00136 \text{ P Propane} \exp(\text{Mm}) + 0.972$$

Gene 5

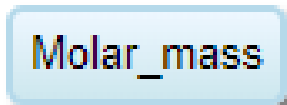
Gene 1



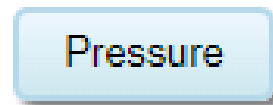
Gene 2



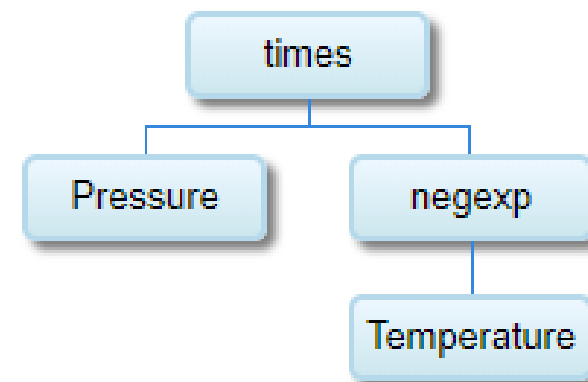
Gene 3



Gene 4



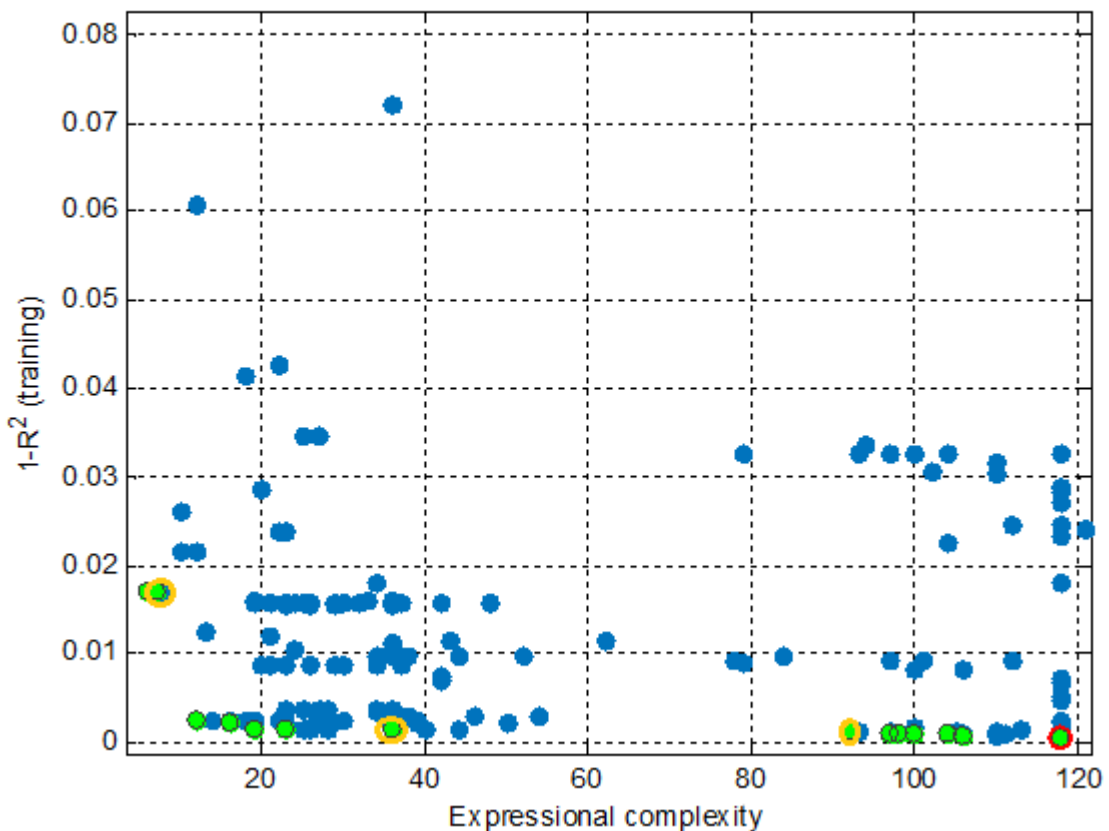
Gene 6





Графики Pareto

Population (merged) models = 900
Data: Z 16 inputs



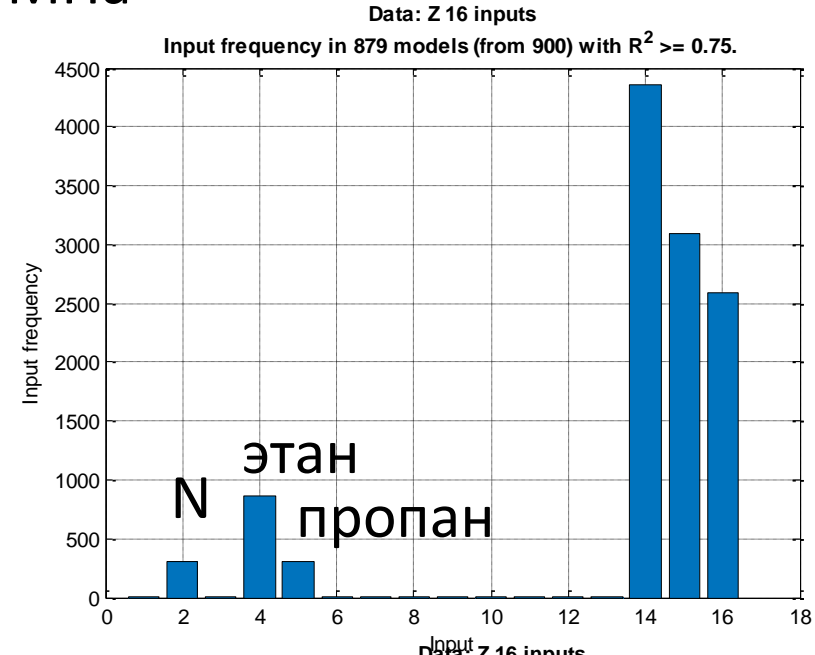
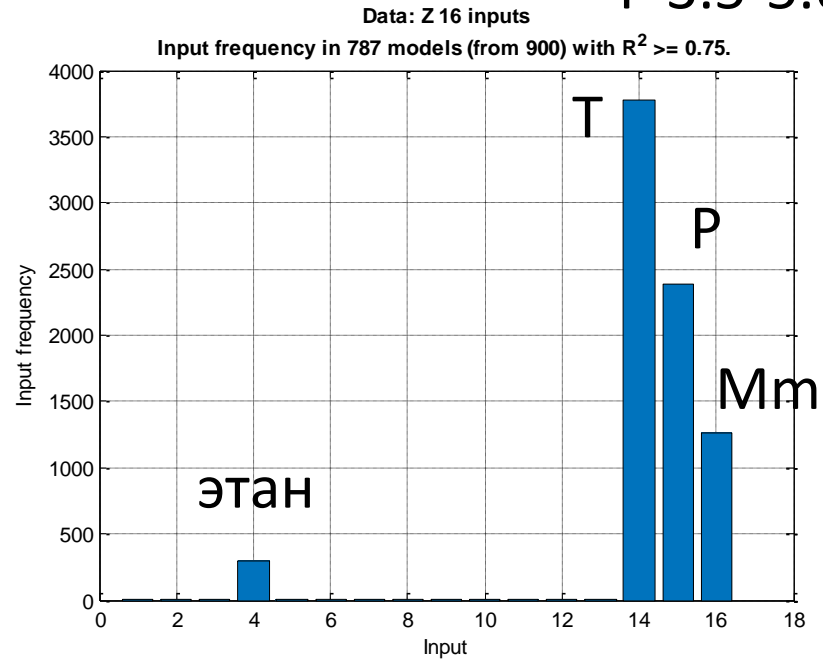
R ²	Model complexity	Model
0.999	118	0.0262 T - 0.0168 P - 0.00804 Mm- 0.00122 T ² (Mm+ P) - 0.00115 (2 Mm+ 2 T) (Mm + P + 2 T) + 0.00872 T (Mm + P) + 0.00114 Mm(CO2 + 2 Methane + N + T) + 0.923
0.999	92	0.0263 T - 0.0169 P - 0.00844 Mm- 0.00134 T ² (Mm+ P) - 0.00105 (2 Mm + 2 T) (Mm + P + 2 T) + 0.00843 T (Mm + P) + 0.923
0.998	36	0.0263 T - 0.0166 P - 0.01 Mm+ 0.00639 P T + 0.00214 T (Mm - P T) - 0.00488 T ² + 0.924
0.983	6	0.0265 T - 0.0176 P - 0.00998 Mm- 0.00478 T ² + 0.924

$$1 - R^2 = SS_{ocmam} / SS_{общ} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 / \sum_{i=1}^n (y_i - \bar{y})^2$$

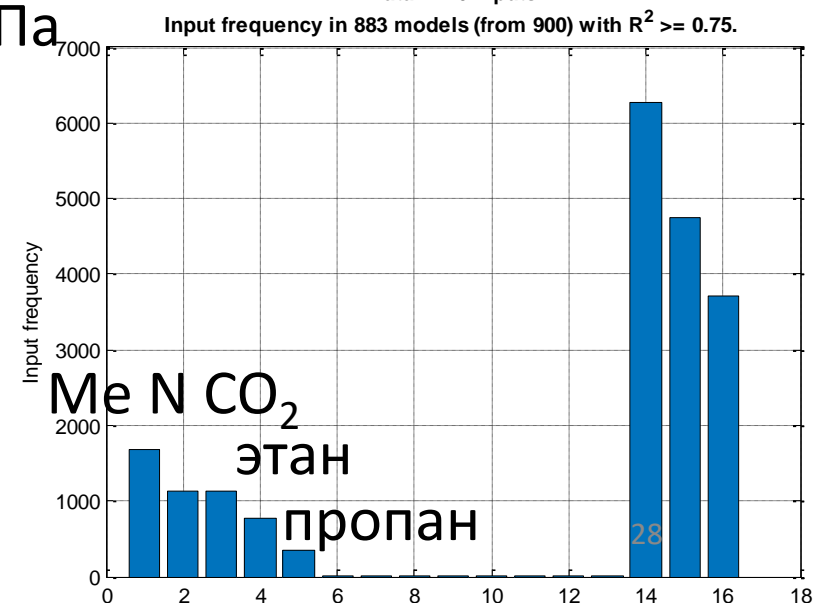
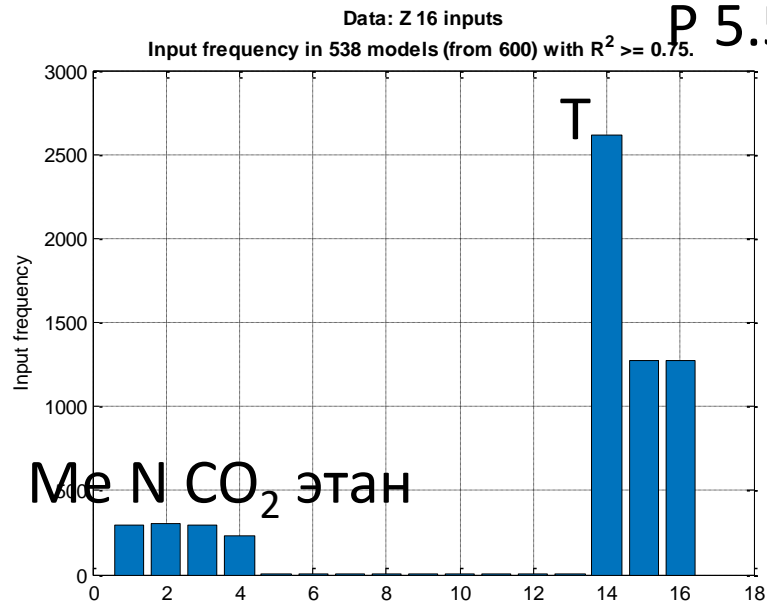
Наиболее значимые входные переменные



Р 3.5-5.6 МПа



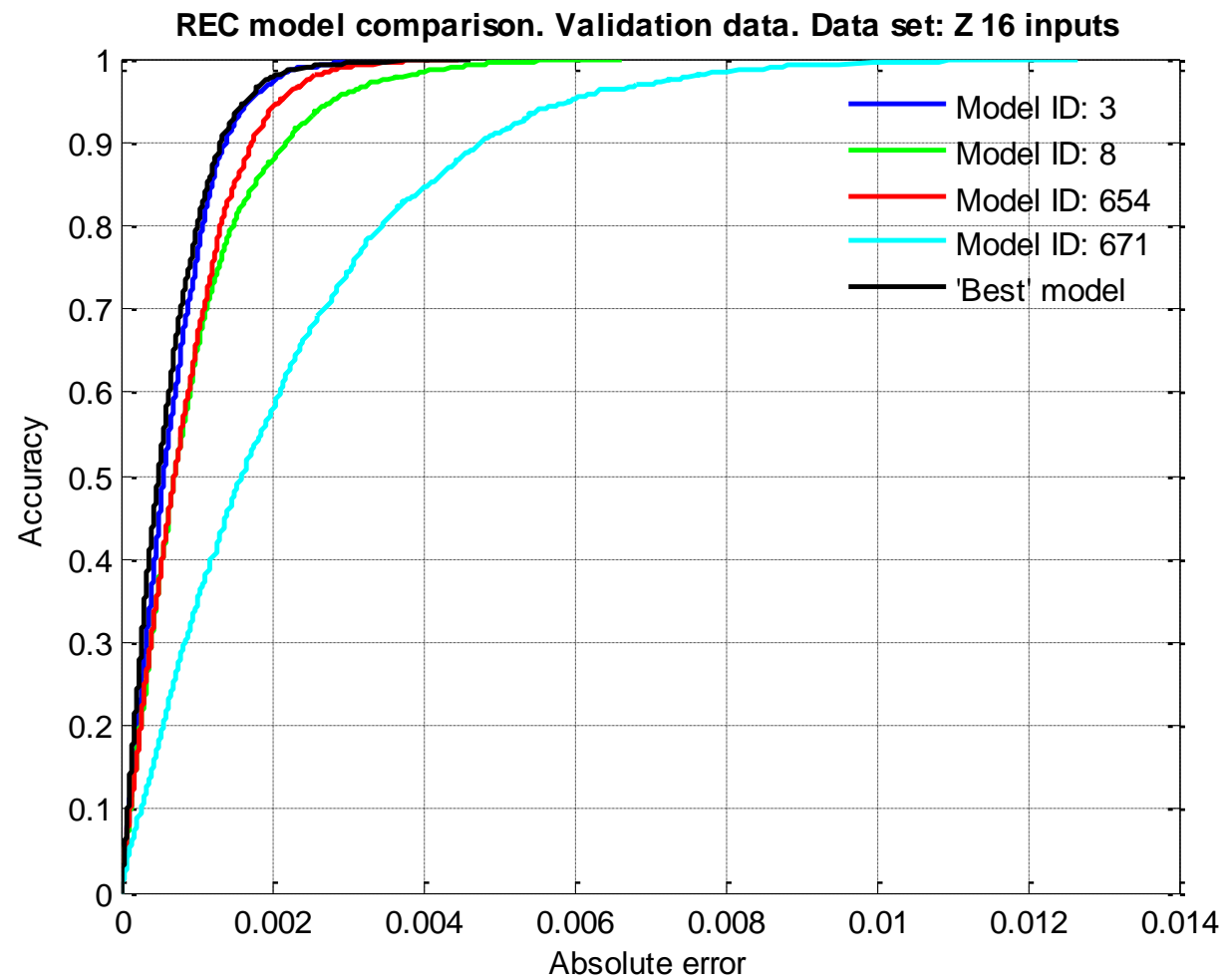
Р 5.5-7.5 МПа





REC – кривые для сравнения моделей

Графики REC (Regression Error Characteristic) строятся для тестовой выборки, по оси абсцисс – модуль отклонения предсказанного значения выходной переменной от реального, по оси ординат – доля данных для которых ошибка предсказания не более x . REC – можно интерпретировать как эмпирическую функцию распределения ошибки построенной модели.



Символьная регрессия или генетическое программирование

Символьная регрессия - метод построения регрессионных моделей путём перебора суперпозиций заранее заданного набора функций, вид функциональной зависимости определяется в процессе работы генетического алгоритма

Достоинства:

Возможно использовать в том случае, когда неизвестен заранее вид модели

Модель интерпретируема – легко поддается анализу экспертов-технологов

Может заменить сложную математическую модель (суррогатное моделирование), в том числе с итерационными процедурами или предполагающую решение дифференциального уравнения разностными методами.

Недостатки:

1. Ресурсоёмкость
2. Нередко полученные модели избыточно сложны