

PB050: Modelování a predikce v systémové biologii

David Šafránek

15.12.2011



Obsah

Parametrizace modelů a estimace parametrů

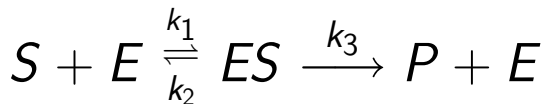
Stochastické vs. deterministické modely

Obsah

Parametrizace modelů a estimace parametrů

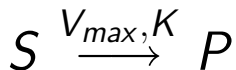
Stochastické vs. deterministické modely

Parametrizace kinetiky enzymů



- model je parametrizovaný množinou parametrů
 $P = \{k_1, k_2, k_3\}$
- bez valuace všech neznámých (neurčitých) parametrů nelze simulovat
- parametry nemusí být přímo experimentálně měřitelné
 - k_1, k_2 prakticky velmi těžko měřitelné *in vitro*
 - k_3 lze získat *in vitro* ale obtížné *in vivo*

Estimace parametrů

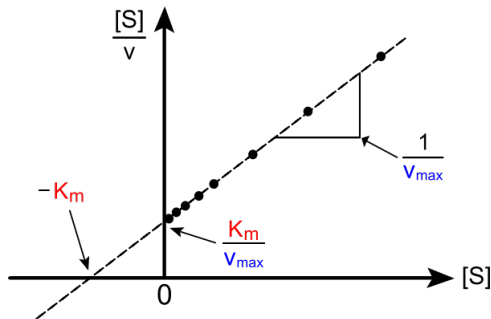


- model je parametrizovaný množinou parametrů
 $P = \{V_{max}, K\}$
- V_{max}, K lze získat nepřímo měřením *in vitro*
 - měří se koncentrace $S(0)$ a odpovídající výkon reakce v
 - po transformaci systému $v = \frac{V_{max}S}{K+S}$ se souřadnicemi v, S na systém v souřadnicích $\frac{S}{v}, S$ dostáváme:

$$\frac{S}{v} = \frac{S}{V_{max}} + \frac{K}{V_{max}}$$

toto lze využít pro lineární regresi (Hanes-Woolf plot)

Estimace parametrů

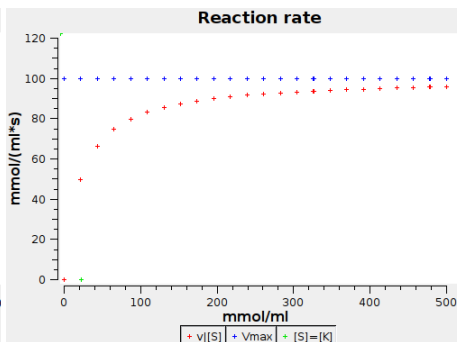
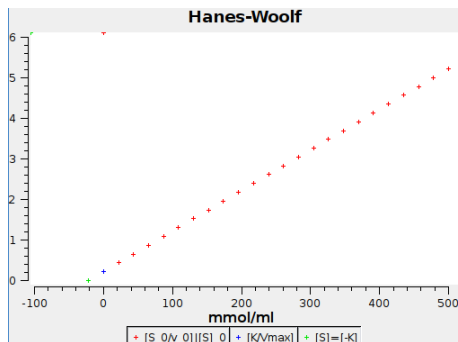


- měří se koncentrace $S(0)$ a iničiální tok v
- lineární regresí získáme K a V_{max}
- problémy: chybovost měření, nerealizovatelné *in vivo*

Hanes, CS (1932). "Studies on plant amylases: The effect of starch concentration upon the velocity of hydrolysis by the amylase of germinated barley.". *Biochemical Journal* 26 (5): 1406-1421.

Estimace parametrů

Demonstrace v COPASI



- závislost iniciálního toku $v(0)$ na iniciální koncentraci $S(0)$
- $S(0)$ samplováno rovnoměrně od 0 do 500
- K (resp. $-K$) znázorněno jako bod na ose S

Estimace parametrů

- co je cílem?
 - najít takovou valuaci parametrů, která nejlépe odpovídá experimentálně zjištěným time-course datům
 - chceme tedy co nejvíce přiblížit simulaci experimentálním datům (tzv. **fitting**)
- lineární regrese požaduje normální rozložení chyb měření
- transformací se nepřesnosti kumulují
- model je inherentně nelineární
 - nelineární jsou i naměřená data
- problém fittingu chápán jako optimalizační problém
- mnoho heuristických metod pro aproximativní řešení
⇒ viz COPASI → Parameter Estimation

Problém reverzního inženýrství

- tzv. **inverzní problémy**
 - cílem je získat model z pozorování systému
 - obecně řešeno v teorii systémů (identifikace systémů)
 - pro nelineární systémy obecně neřešitelné
 - viz. IV120
- obecné schema řešení inverzního problému:
 1. identifikace vztahů mezi proměnnými
 2. identifikace funkcí popisujících sémantiku jednotlivých vztahů (např. zákon zachování hmoty, Michaelis-Menten, Hill, ...)
 3. **estimace hodnot parametrů ve funkcích získaných v předch. bodě**

Estimace parametrů optimalizací

- obecný postup:
 1. srovnej experimentální time-course se simulovaným time-course
 2. pokud rozdíl menší než nastavená tolerance → DONE
jinak modifikuj parametry modelu
 3. proved time-course simulaci modelu
 4. iteruj (1)

Estimace parametrů optimalizací

- mějme model daný systémem $\frac{dx}{dt} = f(x, p)$ kde x je stavový vektor a p je **vektor hodnot parametrů**
- uvažujme $T = \langle t_1, \dots, t_m \rangle$ rostoucí posloupnost časových bodů (tzv. **časovou řadu**)
- předpokládejme posloupnost $\langle x(t_1), \dots, x(t_m) \rangle$ je aproximace řešení $x(t)$ zachycená v časové řadě T (**simulace**)
- zdůrazněme fakt, že simulace byla získána při nastavení hodnot parametrů p , označením $x(t)_p$
- mějme **experiment** jako posloupnost vektorů naměřených veličin $\langle y(t_1), \dots, y(t_m) \rangle$ v časové řadě T
- pro jednoduchost uvažujme $\dim(x) = \dim(y) = 1$ (obecně $\dim(x) \geq \dim(y)$ libovolné, ale složitější formulace)

Estimace parametrů optimalizací

- definujeme odchylku experimentu od simulace v časovém bodě t_i jako tzv. **reziduál**:

$$r(t_i, p) = y(t_i) - x(t_i)_p$$

- reziduál chápeme jako funkci závislou na nastavení parametrů simulovaného modelu
- srovnání experimentu a simulace je vyjádřeno jako součet čtverců reziduálů přes vš. časové body T :

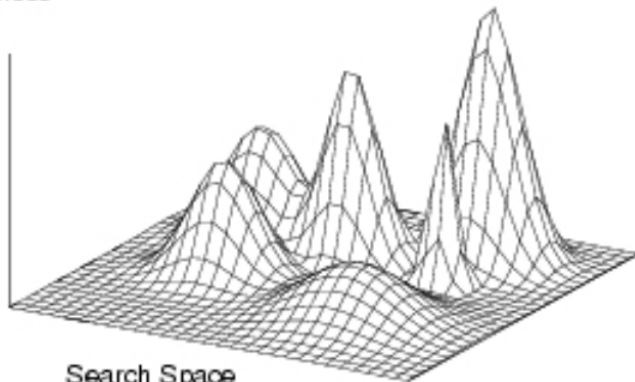
$$S(p) = \sum_i^m (r(t_i, p))^2$$

Estimace parametrů optimalizací

- funkce $\mathcal{S}(p)$ se nazývá **užitková funkce**
- vystihuje průměrnou odchylku simulace od experimentu přes danou časovou řadu
- **minimální hodnota** $\mathcal{S}(p)$ určuje optimální vektor hodnot parametrů p , který globálně minimalizuje rozdíl mezi experimentem a modelem
- jedná se o nelineární funkci
- počet neurčitých parametrů určuje její dimenzi

Optimalizační krajina

Fitness



Procházky po optimalizační krajině...

- CÍL: najít **globální minimum**
- nejpoužívanější jsou stochastické black-box přístupy:
 - náhodné procházení (random search)
 - evoluční strategie (evolution strategy)
 - ...
- black-box znamená absolutní nezávislost na tvaru užitkové funkce
- existují i metody, které využívají znalosti užitkové funkce (např. simulované žíhání, Truncated Newton, ...)

Procházky po optimalizační krajině...

Random Search

1. inicializuj náhodně výchozí hodnotu p
 - typicky z rovnoměrného rozložení
2. dokud není překročen povolený počet iterací, prováděj:
 - 2.1 sampluj novou pozici p'
 - uniformní náhodný výběr z hyperkoule o daném poloměru
 - 2.2 spočítej $\mathcal{S}(p')$
 - 2.3 pokud $\mathcal{S}(p') < \mathcal{S}(p)$, nastav novou pozici $p := p'$
3. p nastaveno na nejvýhodnější pozici (z pohledu běhu algoritmu)

Pozn. Existují varianty s fixním i adaptivním poloměrem.

Procházky po optimalizační krajině...

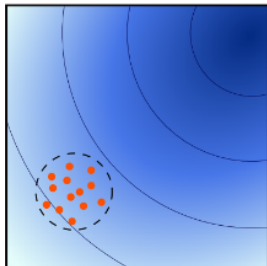
Evoluční strategie

- inherentně adaptivní metoda, redukuje počet
- staví na výběrech z (vícerozměrného) normálního rozložení
- rozměr daný dimenzí vektoru parametrů
- značeno CMA-ES (Covariance Matrix Evolution Strategy)
- postup:
 1. vytvoř generaci
 - sampluj rozložení pozic hodnot P dle normálního rozložení
 2. pro každé $p \in P$ spočítej $S(p)$
 3. adaptace: uprav parametry normálního rozložení pro další iteraci
 - různé varianty, mohou být velice komplexní adaptace citlivé na charakter evoluční krajiny – adaptace kovarianční matice
- metoda se ukazuje výhodná pro biologické modely (vysoká míra neznalosti parametrů)

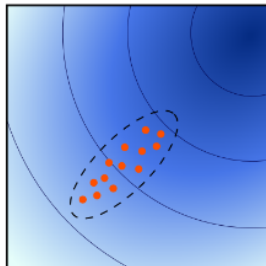
Procházky po optimalizační krajině...

Evoluční strategie

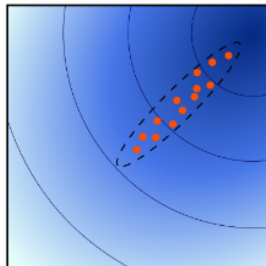
First generation



Second generation



Third generation



Tuning dynamických modelů

Estimace parametrů vzhledem k hypotézám

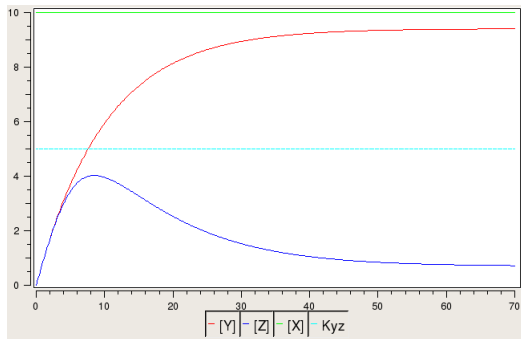
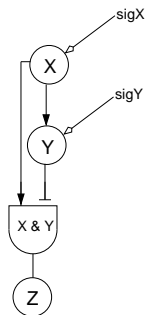
- estimaci parametrů lze rozšířit pro temporální specifikace (hypotézy)
 - každý experiment lze zakódovat jako LTL formuli
 - typické abstraktnější vlastnosti dávají větší stupeň volnost
- mějme specifikaci chování reprezentovanou LTL formulí φ
- uvažme parametrizovaný systém $\frac{dx}{dt} = f(x, p)$
- uvažme časovou řadu T
- $x_{p,T}$ značí simulovanou posloupnost v časové řadě T s nastavením parametrů p
- definujme uživatelskou funkci

$$S(p) = vd(x_{p,T}, \varphi)$$

Tuning dynamických modelů

Příklad I – BioCHAM

- uvažujme model genového regulačního obvodu IFFL



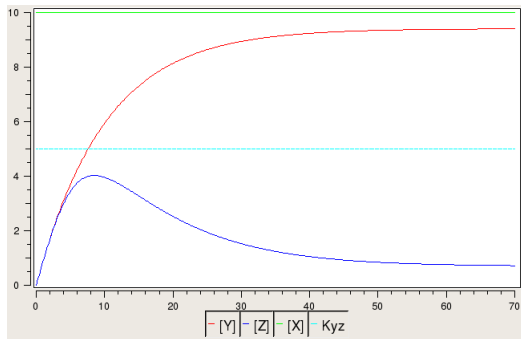
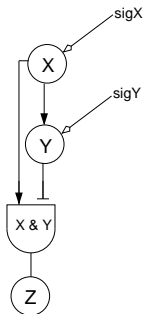
$$\beta_y = \beta_z = 1, \gamma_y = \gamma_z = 0.1, X = 10 \text{ (konst.)}$$

$$n = 4, K_{xy} = 4, K_{xz} = K_{yz} = 5$$

Tuning dynamických modelů

Příklad I – BioCHAM

- uvažujme model genového regulačního obvodu IFFL

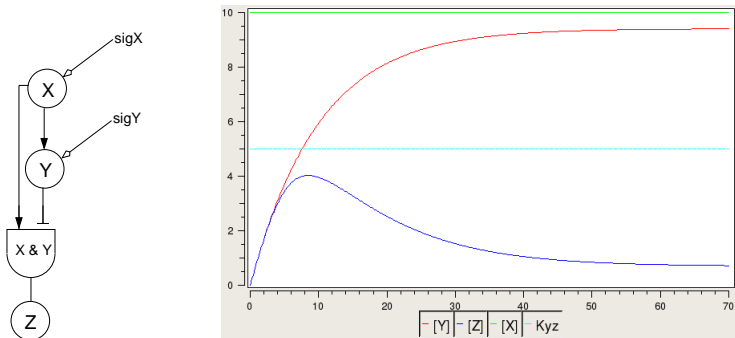


Pulsní chování Z splňuje $\mathbf{F}([Z] \geq 3)$.

Tuning dynamických modelů

Příklad I – BioCHAM

- uvažujme model genového regulačního obvodu IFFL

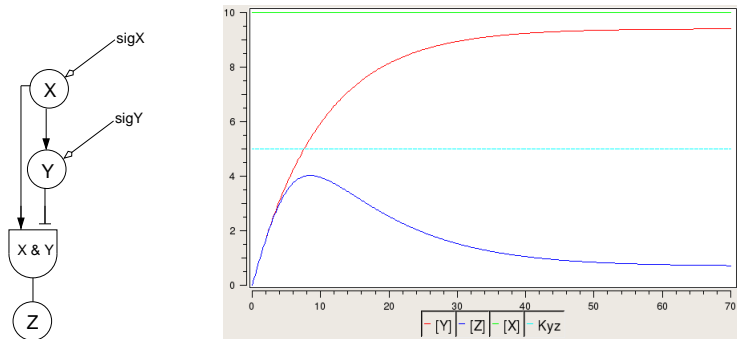


Navrhňte tuning parametru β_z tak, aby puls přesáhl 5.

Tuning dynamických modelů

Příklad I – BioCHAM

- uvažujme model genového regulačního obvodu IFFL

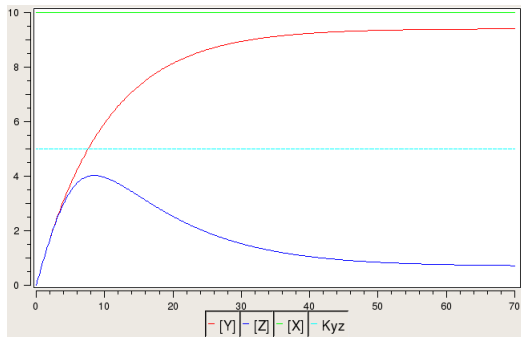
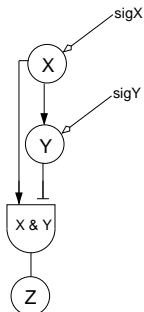


Uplatníme vlastnost $\mathbf{F}([Z] \geq 5)$.

Tuning dynamických modelů

Příklad I – BioCHAM

- uvažujme model genového regulačního obvodu IFFL



BioCHAM:

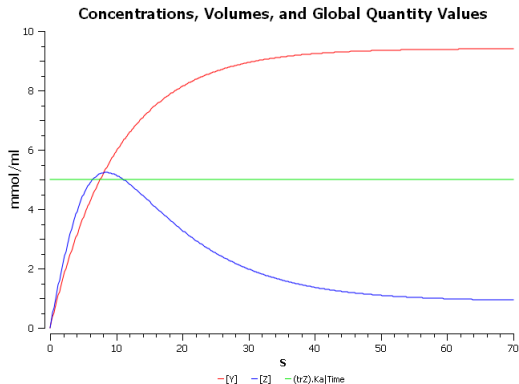
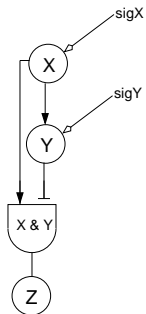
```
search_parameters_cmaes([ $\beta_z$ ], [(0, 5)], F([Z] >= v), [v], [5], 100)).
```

– nalezení hodnoty β_z v rozsahu [0, 5] pro simulaci do 100s

Tuning dynamických modelů

Příklad I – BioCHAM

- uvažujme model genového regulačního obvodu IFFL

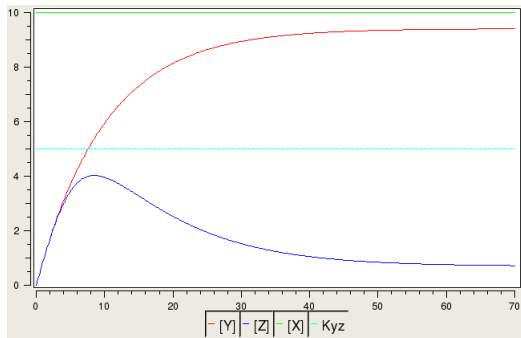
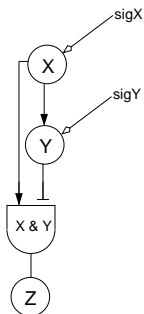


Výsledek je $\beta_Z = 1.3$.

Tuning dynamických modelů

Příklad II – BioCHAM

- uvažujme model genového regulačního obvodu IFFL

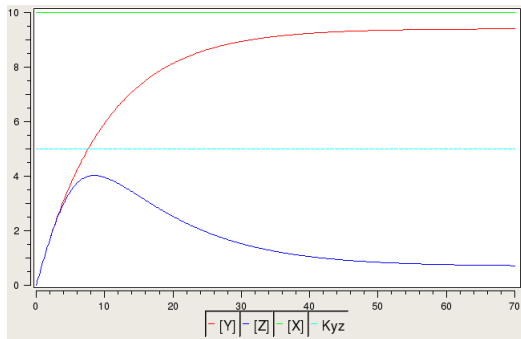
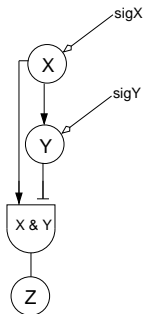


Navrhněte tuning parametrů β_z, γ_z tak, aby puls přesáhl 5 do 5s.

Tuning dynamických modelů

Příklad II – BioCHAM

- uvažujme model genového regulačního obvodu IFFL

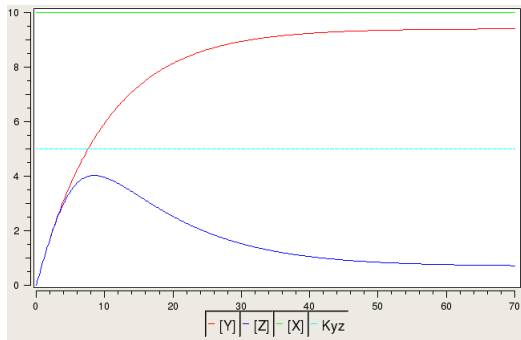
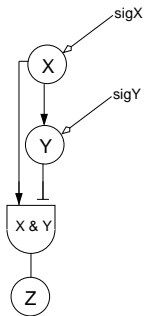


Uplatníme vlastnost $\mathbf{F}([Z] \geq 5 \& \text{Time} \leq 5)$.

Tuning dynamických modelů

Příklad II – BioCHAM

- uvažujme model genového regulačního obvodu IFFL



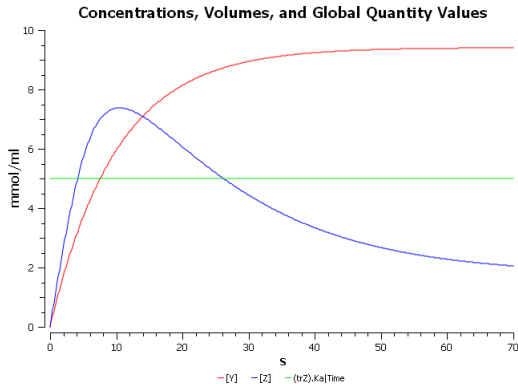
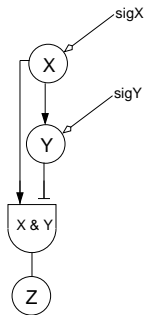
BioCHAM:

```
search_parameters_cmaes([\beta_z, \gamma_z], [(0, 5), (0, 5)], F([Z] >= v)
&(Time = < 5)), [v], [5], 100)).
```

Tuning dynamických modelů

Příklad II – BioCHAM

- uvažujme model genového regulačního obvodu IFFL

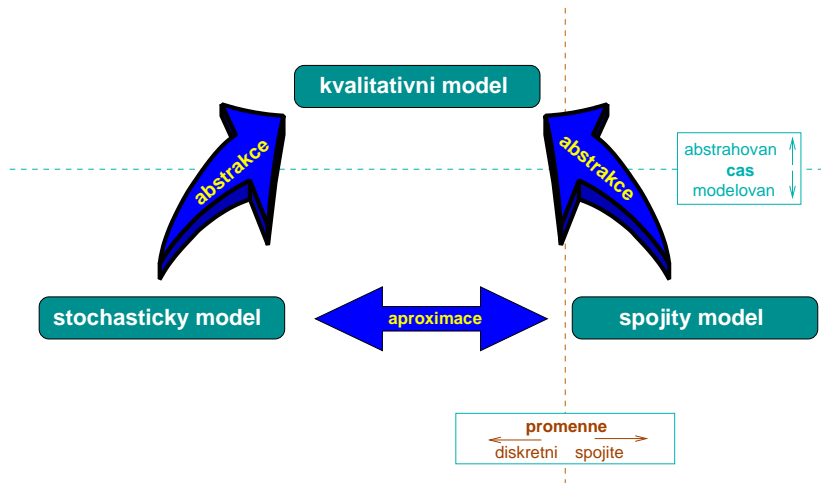


Výsledek je $\beta_Z = 1.498$, $\gamma_Z = 0.059$.

Obsah

Parametrizace modelů a estimace parametrů

Stochastické vs. deterministické modely

Spojité vs. diskrétní model

Stochastický vs. deterministický model

- molární koncentrace $[M]$:

$$C = \frac{n}{V}$$

kde n je množství látky $[mol]$, V je objem roztoku $[l]$

- vyjadřuje se pomocí Avogadrovy konstanty (počet částic v 1 molu):

$$C = \frac{N}{N_A \cdot V}$$

kde N_A Avogadrova konstanta $[mol^{-1}]$, V objem roztoku $[l]$ a N je počet molekul.

- převodní faktor $\gamma = N_A \cdot V$:

$$N = C \cdot \gamma$$

Příklad

- bakterie *E. coli* je válcový objekt o délce $v \sim 2\mu\text{m}$ a poloměru $r \sim 1\mu\text{m}$
- objem lze tedy odhadnout následujícím způsobem:

$$V = \pi r^2 v = \pi(0,5 \cdot 10^{-6})^2 \cdot (2 \cdot 10^{-6}) = \frac{\pi}{2} \cdot 10^{-18} \text{m}^3 = \frac{\pi}{2} \cdot 10^{-15} \text{L}$$

- pokud uvažujeme látku X s koncentrací $[X] = 10^{-5} \text{M}$ rozpuštěnou v takovéto buňce *E. coli*, počet molekul $\#X$ je:

$$\#X = [X]N_A V = 10^{-5} \cdot 6,023 \cdot 10^{23} \cdot \frac{\pi}{2} \cdot 10^{-15} = 9461$$

Stochastický vs. deterministický model

Pro reakci R_i definujeme převodní vztah mezi stochastickou a deterministickou variantou zákona o zachování hmoty tím, že dáme do souvislosti stoch. frekvenci c_i a det. kinetickou konstantou k_i :

typ reakce R_i	$c_i \rightarrow k_i$
$S_j \rightarrow *$	$k_i = c_i$
$S_p + S_q \rightarrow *$	$k_i = c_i \cdot \gamma$
$2S_j \rightarrow *$	$k_i = \frac{c_i \cdot \gamma}{2}$

Stochastický vs. deterministický model

Pro reakci R_i definujeme převodní vztah mezi stochastickou a deterministickou variantou zákona o zachování hmoty tím, že dáme do souvislosti stoch. frekvenci c_i a det. kinetickou konstantou k_i :

typ reakce R_i	$c_i \rightarrow k_i$
$S_j \rightarrow *$	$k_i = c_i$
$S_p + S_q \rightarrow *$	$k_i = c_i \cdot \gamma$
$2S_j \rightarrow *$	$k_i = \frac{c_i \cdot \gamma}{2}$

Pozn.: Složitější kinetiky odvozené na základě kvasistabilní abstrakce nelze přímo převádět.