# A Hybrid HMM Approach for the Dynamics of DNA Methylation

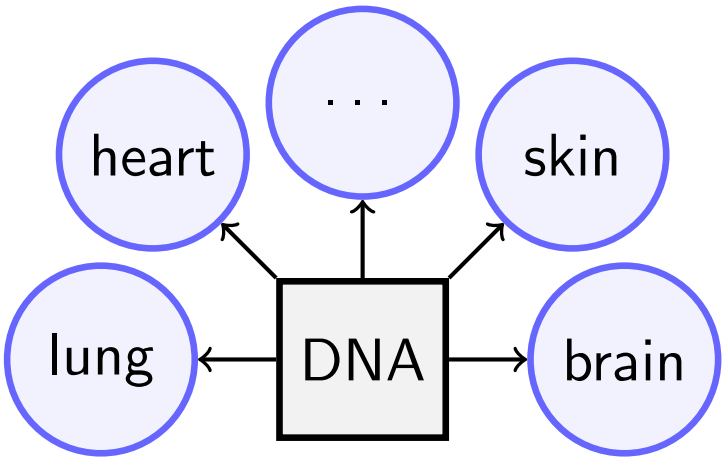
**Charalampos Kyriakopoulos, Pascal Giehr, Alexander Lück, Jörn Walter, Verena Wolf**

UNIVERSITÄT
DES
SAARLANDES

HSB 2019 6th Workshop on Hybrid Systems & Biology
April 06, 2019

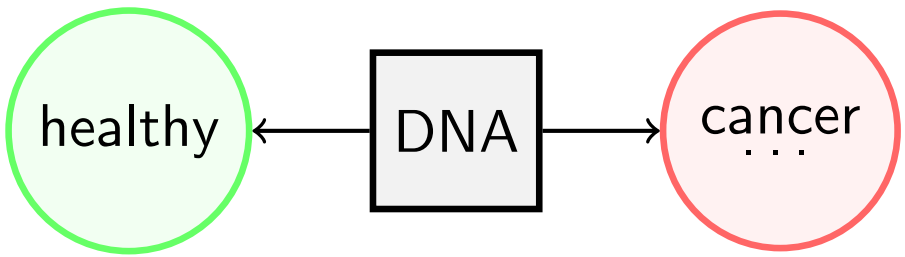
# Importance of Epigenetics

**Every cell contains the whole genome and therefore the "blueprints" for all producible proteins.**




How can cells specialize?

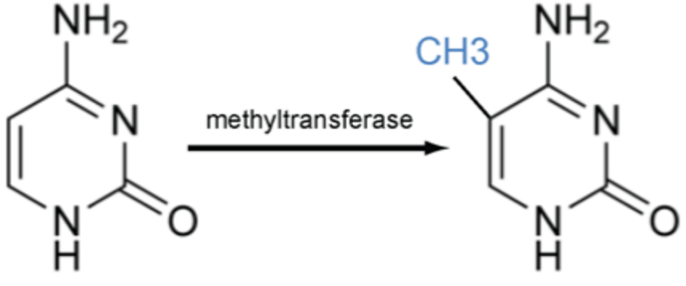
Why do some cells develop diseases?



# DNA Methylation

...T A C G C C C T G T C G A...
...A T G C G G G A C A G C T...

Occurs (almost exclusively) on **cytosine** in **CpGs** (DNA sequence: C - Phosphor - G)




**C** **5mC**

DNA methyltransferases (Dnmts) convert cytosine (C) to 5-methylcytosine (5mC)

**Methylated cytosine hinders transcription of DNA into mRNA**

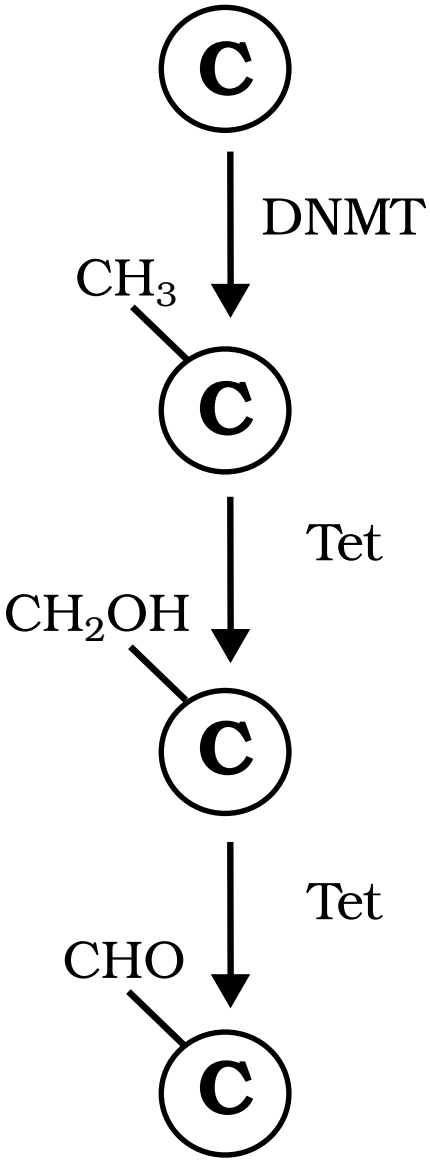
# Methylation and further modifications

cytosine (C):
original unmodified base in DNA

5-methylcytosine (5mC):
methylated C $\rightarrow$ gene inactivation

5-hydroxymethylcytosine (5hmC):
hydroxymethylated C $\rightarrow$ gene activation

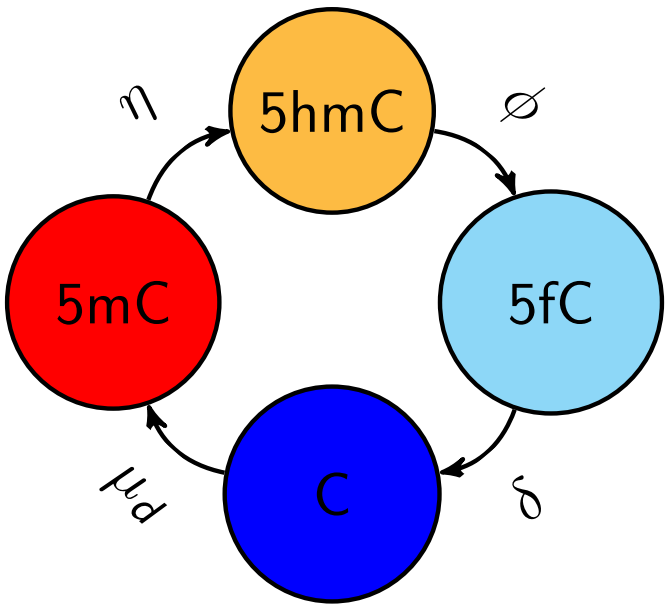
5-formylcytosine (5fC):
formylated C $\rightarrow$ active demethylation



# Passive and active demethylation

**Passive demethylation:**
Losing methylation over time due to cell division and failed maintenance and/or decreasing methylation efficiency.

**Active demethylation:**



# Motivation

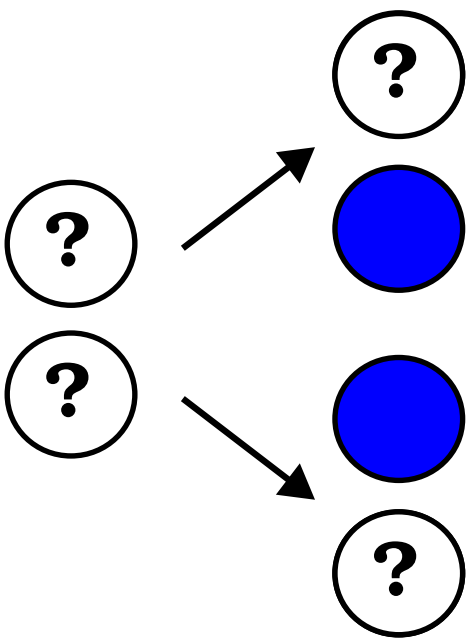
1. Incorporate different types of event classes in the model: Events that only occur once at deterministic times (*discrete*; cell division and maintenance) and events that may occur more than once at random times (*continuous*; methylation cycle). $\rightarrow$ Existing models are either exclusively discrete or continuous.

2. Model and predict levels of 5fC.
   $\rightarrow$ Important for active demethylation.

# Modeling of (de)methylation events

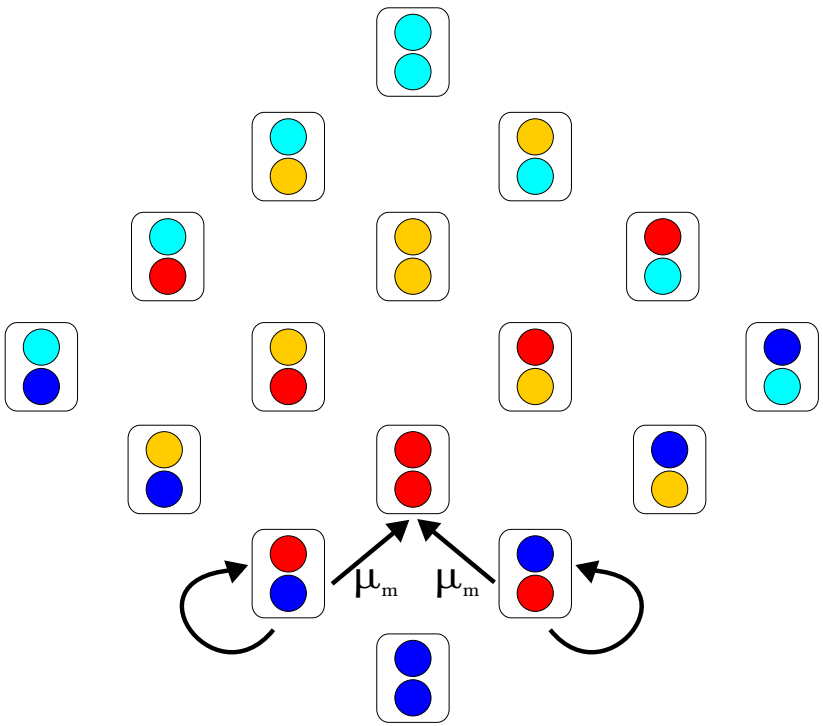
Modeling via **Hidden Markov Model** with 16 hidden states (4 methylation states on double-stranded DNA) and 4 observable states (2 for each strand).

A detailed look into the individual processes:




**Cell division:**
Keep one strand as it is and synthesize a new complementary strand with only unmethylated cytosines.
Strand to keep is chosen randomly with probability 0.5.

# Maintenance methylation

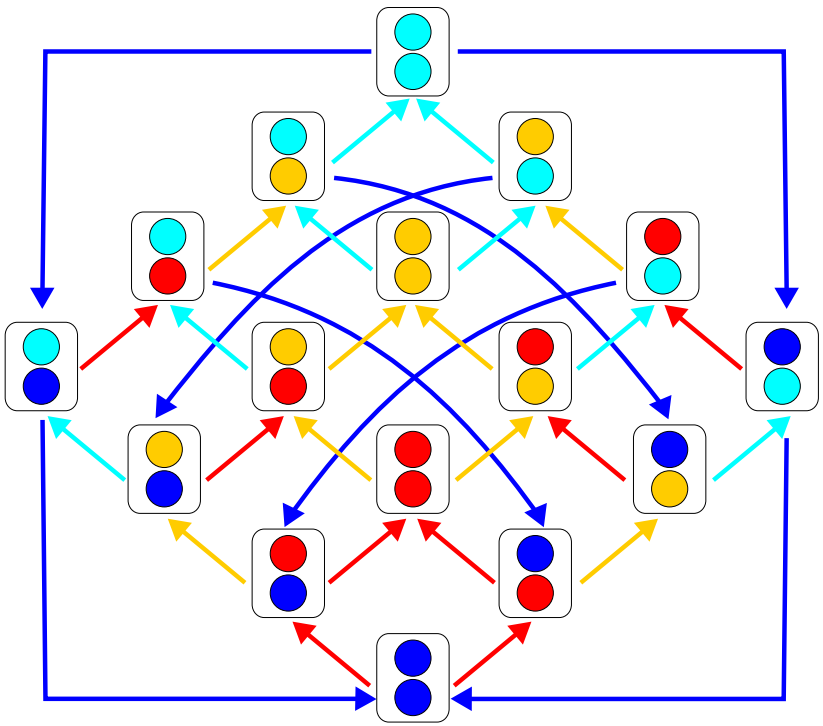



States: C, 5mC, 5hmC, 5fC

Maintenance methylation only on hemi-methylated CpGs (rate $\mu_m$) right after cell division.

Assumption based on previous results:
No maintenance on hemi-hydroxylated and hemi-formylated CpGs

# Continuous transitions



States: C, 5mC, 5hmC, 5fC

The reactions

*de novo* $\mu_d$

hydroxylation $\eta$

formylation $\phi$

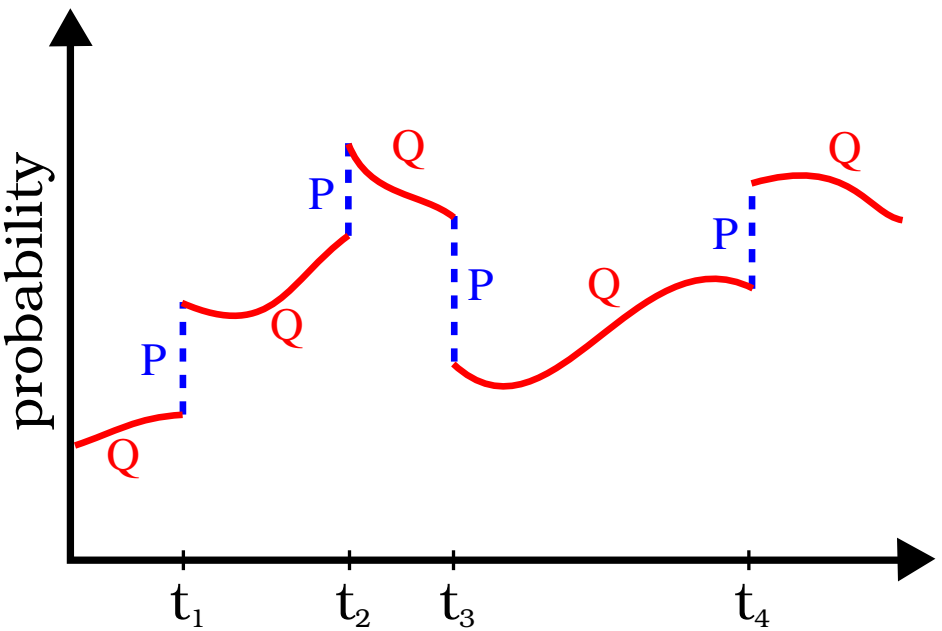
demethylation $\delta$

may occur more than once in each cell division cycle.

# Model dynamics

**Discrete transitions** at *fixed* times $t \in \{t_1, t_2, \ldots, t_n\}$:
Cell division and maintenance (linked to replication fork)
$\rightarrow$ DTMC with transition probability matrix **P**

**Continuous transitions** at *random* times $t \in [t_i, t_{i+1}]$:
Transitions within the methylation cycle (*de novo*, hydroxylation, formylation, active demethylation)
$\rightarrow$ CTMC with infinitesimal generator matrix **Q**



# Efficiencies

Let $r \in \{\mu_m, \mu_d, \eta, \phi, \delta\}$, where $\mu_m$ is a transition probability and $\mu_d,\ \eta,\ \phi,\ \delta$ are transition rates.

Time dependent efficiencies, due to e.g. changing enzyme concentrations:

$$r(t) := \alpha_r + \beta_r \cdot t$$

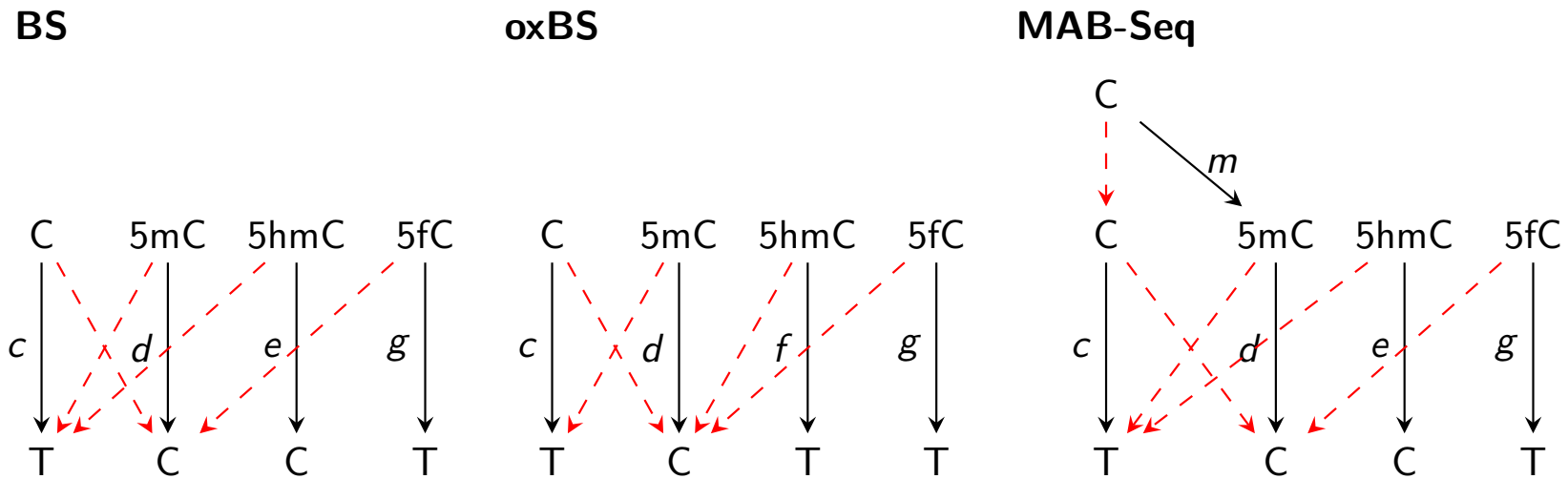
Introduce bounds in order to ensure *identifiability*:

$$0 \leq r(t) \leq ub$$

Choose *ub* based on biological assumptions
$\rightarrow$ prohibit arbitrarily fast reactions

# Hidden and observable states

**Hidden states:** C, 5mC, 5hmC and 5fC
**Observable states:** C and T

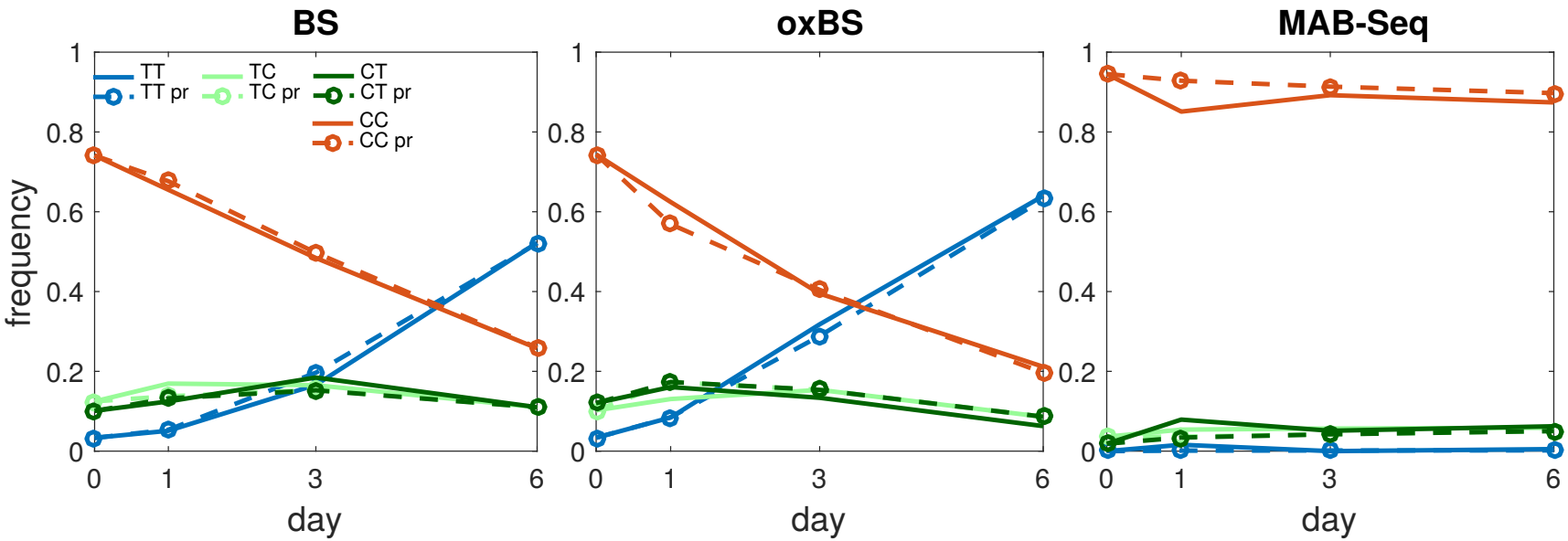


# Data

**Example data set:**
Data for single copy gene Afp (alpha fetoprotein) containing 5 CpGs. Three data sets (BS, oxBS and MAB-Seq) to identify hidden states.

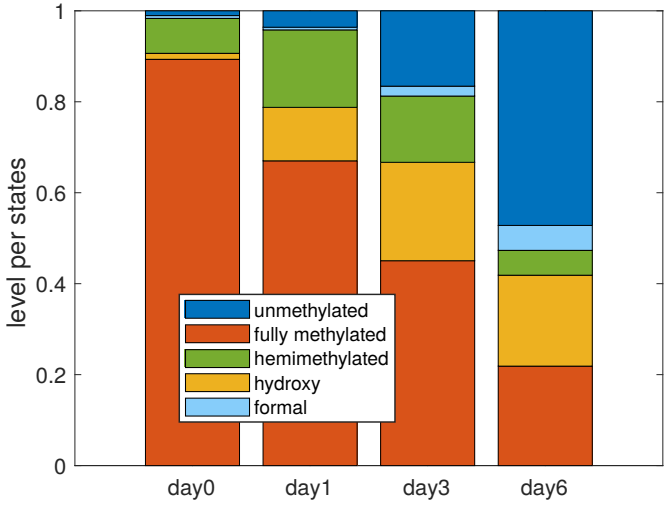
**Parameter estimation** via MLE. Similar results for all 5 CpGs
$\Rightarrow$ Show only aggregated results in the following.

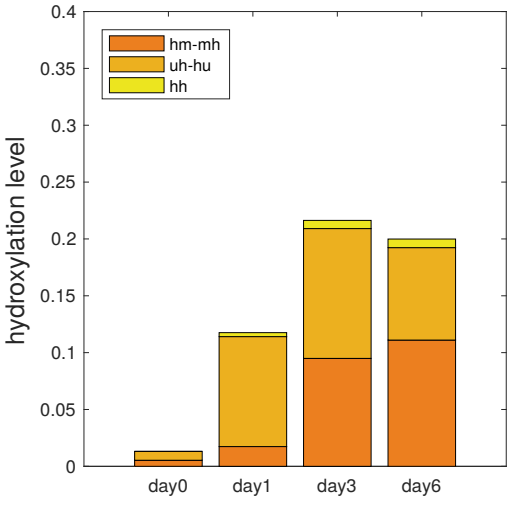
# Observable states

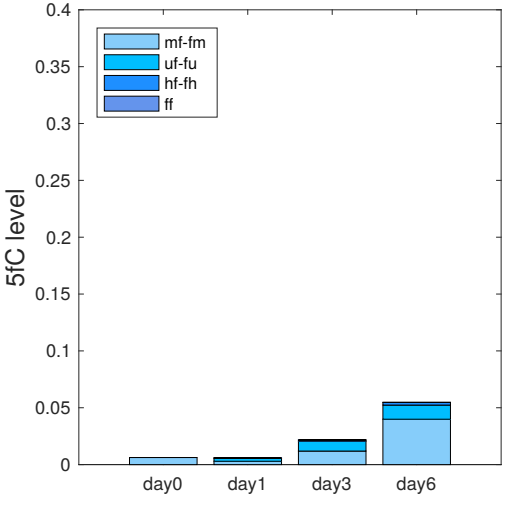



Good agreement between data (solid lines) and results from the model (dashed lines).

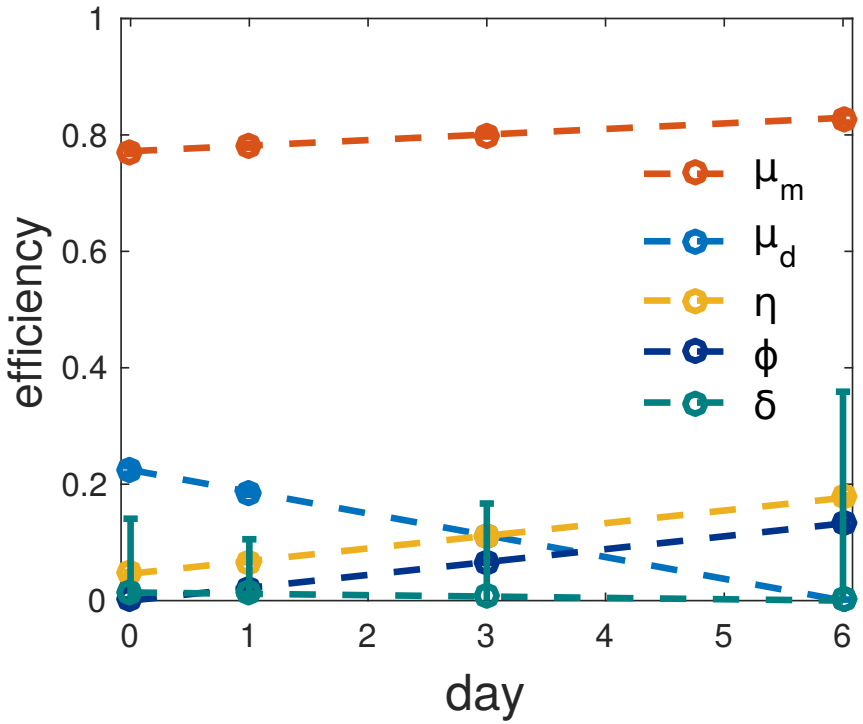
# Hidden states



# Efficiencies




Measured time points too far apart (only one measurement for each cell division cycle) ⇒ How often was the methylation cycle traversed?

No information about intermediate time points.

# Summary

- Hybrid HMM (discrete for cell division and maintenance; continuous for methylation cycle events)
- Very flexible model (choice of time points for discrete events, efficiency function)
- Good prediction performance, however available data is not ideal so far

# Thank you for your attention!