Formal Executable Descriptions of Biological Systems

Pierpaolo Degano

Dipartimento di Informatica,
Università di Pisa, Italia

joint work with a lot of nice people :-)

Pisa, 14th June 2007

From Syntax to Semantics

To understand function, study structure – F. Crick

seems to work no longer in modern biology:

STRUCTURE ***AND*** FUNCTION

The genome as a 4-letters language — syntax

⇓

what and how it expresses for — semantics

Systems Biology (a partial view)

- Hypothesis-driven investigation in place of reductionism
  - build a formal model of a biological system (generation of hypothesis)
  - experiment it (tuning of hypothesis) until the model gets validated and ready to use
- Leads to a global view of a system — but often only offers snapshots of its behaviour
- Huge amount of data available — hard to handle, very hard to interpret

Computer Science (similarities)

- A computer systems is
  - formally modelled (generation of hypothesis)
  - implemented, refined and eventually validated (experimenting on hypothesis)
- Experiments requires executing the model, to obtain its whole behaviour
- Analysis methods and tools exist

... and computational power increasingly grows

Long term goals

- Understand the functionality of bio-components
  - assessment of known facts
  - discovery of new functionalities
- Investigate the underlying structure of biological complex systems
  - how genome, proteome and metabolome interact giving rise to *emergent properties*

Mathematical description of bio-phenomena

- bio-physics – since Schrödinger, lots of differential equations, with deep statistical and stochastic models (monolithic, large, difficult to state, change, adapt and ... to solve for me:-)

Mathematical description of bio-phenomena

- bio-physics – since Schrödinger, lots of differential equations, with deep statistical and stochastic models (monolithic, large, difficult to state, change, adapt and ... to solve for me:-)
- bio-informatics:
  – structure (human) genome (DNA as a formal language over ACGT) and data bases of genes, proteins, metabolic pathways, ...

Mathematical description of bio-phenomena

- bio-physics – since Schrödinger, lots of differential equations, with deep statistical and stochastic models (monolithic, large, difficult to state, change, adapt and ... to solve for me:-)
- bio-informatics:
  – structure (human) genome (DNA as a formal language over ACGT) and data bases of genes, proteins, metabolic pathways, ...
  – function Petri nets, Process calculi, Rewriting systems, ...

"cells *as* computational devices"

Bio-systems

Metabolic and gene regulation networks, signalling pathways, etc are made of

Bio-systems

Metabolic and gene regulation networks, signalling pathways, etc are made of

- millions of components acting independently, interacting each other, dispersed in solutions

Bio-systems

Metabolic and gene regulation networks, signalling pathways, etc are made of

- millions of components acting independently, interacting each other, dispersed in solutions
- interaction
  - is essentially binary

Bio-systems

Metabolic and gene regulation networks, signalling pathways, etc are made of

- millions of components acting independently, interacting each other, dispersed in solutions
- interaction
  - is essentially binary
  - occurs on selected sites (if any) between close enough, affine, non-separated components

Bio-systems

Metabolic and gene regulation networks, signalling pathways, etc are made of

- millions of components acting independently, interacting each other, dispersed in solutions
- interaction
  - is essentially binary
  - occurs on selected sites (if any) between close enough, affine, non-separated components
  - is local, but affects the whole system globally

Bio-systems

Metabolic and gene regulation networks, signalling pathways, etc are made of

- millions of components acting independently, interacting each other, dispersed in solutions
- interaction
  - is essentially binary
  - occurs on selected sites (if any) between close enough, affine, non-separated components
  - is local, but affects the whole system globally

Just as concurrent, distributed, mobile processes

Processes

Concurrent, distributed, mobile processes are made of

- several components acting independently, interacting each other, distributed geographically
- interaction
  - is mainly binary
  - occurs on selected channels between components
  - is local, but affects the whole system globally

Process calculi: primitives

Few basic primitives for

- sending $!a(v)$ and receiving $?a(v)$ the value v, if any, on channel a
  channels mimick *interaction* points, values the exchanged *information*
- performing non detailed activities τ
  abstracting from, e.g., biochemical details
- creating/handling channels

composed with few operators ...

Process calculi: composition

Among the few operators there are:

- parallel composition $P \mid Q$
  cells as processes, that may interact or proceed independently
- choice $P + Q$
  according to a *probabilistic distribution* — more to come

Process calculi: semantics

How do systems evolve?

- Semantics is given through a logically based inference system, defining transitions — how a configuration changes into another
- Communication, i.e. interaction, is the basic computational step

Process calculi: Semantics

Essentially, communication and asynchrony are ruled by:

- $?a(x).P \mid !a(v).Q \rightarrow P[x \mapsto v] \mid Q$

  the activity is local

- IF $P \rightarrow P'$ THEN $P \mid Q \rightarrow P' \mid Q$

  its effect is global — more to come

Quantitative information

... otherwise "*stamp collection*" — Rutherford

- interactions occur at given rates – channels posses rates
- (often) interactions are reversible (possibly with different rates)
- the context affects the overall rates – not only temperature, pressure, etc, but also concentration – here the quantities of reactants per unit (typically, Gillespie's Stochastic Simulation Algorithm)

Summing up

- molecules, metabolites, compounds, cells as processes
- (biochemical) interactions as communications
- affinity of interaction as communication capabilities

(other features, like membranes, geometry, time, ... often treated *ad hoc* or still under investigation)

Process calculi specify and execute Bio-systems

What do we gain?

- run the model, and obtain virtual experiments — an integral abstract description of system behaviour: unexpected, global properties may emerge
- formally analyse the executions, collecting e.g. statistical data on behaviour, or causality among interactions, or similarities/differences between systems, ...
- compositionality — specify new components in isolation (e.g. active principles), put them aside the others with *no other change* and see (cf. ODE)

A simple example

Consider the enzyme-catalysed production of a product P from the substrate S:

$$E + S \rightleftharpoons^{K_{ES}}_{K_{ES}^{-1}} ES \multimap^{K_P} E + P$$

The corresponding processes are

$E =!a$ where $rate(a) = K_{ES}$

$S =?a.ES$ where $rate(\tau_1) = K_{ES}^{-1}$

$ES = \tau_1.(E|P) + \tau_{-1}.(E|S)$ where $rate(\tau_{-1}) = K_P$

A computation is

$E = !a$ where $rate(a) = K_{ES}$

$S = ?a.ES$ where $rate(\tau_1) = K_{ES}^{-1}$

$ES = \tau_1.(E|P) + \tau_{-1}.(E|S)$ where $rate(\tau_{-1}) = K_P$

$$n \cdot E \mid m \cdot S \xrightarrow{r_0}$$

$$(n-1) \cdot E \mid (m-1) \cdot S \mid ES \xrightarrow{r_0'}$$

$$(n-2) \cdot E \mid (m-2) \cdot S \mid 2 \cdot ES \xrightarrow{r_1}$$

$$(n-1) \cdot E \mid (m-2) \cdot S \mid ES \mid P \xrightarrow{r_0''}$$

$$(n-2) \cdot E \mid (m-3) \cdot S \mid 2 \cdot ES \mid P \xrightarrow{r} \ldots$$

where the actual rates $r_0, r_0', \ldots$ are typically computed with Gillespie's SSA and depend on the rates of channels and on the number of reactants.

Other approaches

- Petri nets
- formal languages (P systems, ...)
- rewriting systems (κ-calculus, calculus of looping sequences, ...)
- logically based formalisms (Pathway logic, ...)
- ...

Our own work

A brief report on two ongoing investigations:

- VIrtual CEll:
  artificial ur-cell, from a simplified prokaryote
  — with a variant of the π-calculus
- E. Coli:
  the whole metabolic pathways, with knock-outs
  — with a very fast (subset of) the π-calculus

Towards a holistic model of a ***whole*** cell: all interactions among metabolic pathways (properties emerge), the whole movie not only snapshots

Building up VICE: the genome

Problems:

- not an arbitrary list of genes
- *small* enough for the sake of computability

Our choice: The "Minimal Gene Set"

- from *Haemophylus influenzae, Mycoplasma genitalium*
- cf. Glass et al. – gene KO *in vitro*

Building up VICE: hypothesis

Reduction and update of the *Minimal Gene Set*, based on a functional analysis.

- selection of basic activities (*eating*, production of energy, synthesis of basic structural components, reproduction)
- choice of the 187 genes involved
- design of the metabolic pathways needed (presently only for *survival*)

VICE: Validation

- Check on biological consistency:
  - all the pathways selected have been taken: *sufficient*
  - no genes are left inactive: *necessary*
- Comparison with real results:
  - confirm basic modelling choice
  - calls for deeper analysis and more features

Activities

Group pathway (and reactions) in the standard biochemical manner:

Oxidations: extraction of energy from nutrients:

Glycolysis$\rightarrow$Pyruvate$\rightarrow$...

Lipid metabolism: synthesis of structural components from monomers: fatty acids...

Nucleotide metabolism: building DNA/RNA bases, no *de novo* synthesis

DNA/RNA synthesys: RNA for building proteins, DNA for reproduction – not yet available

Protein synthesis: no amino acids

Uptake: Glycerol, amino acids, nitrous bases, fatty acids...

...plus a few other pathways.

Virtual experiments

Through runs of the π-specification of VICE

- in presence of different quantities of food (VICE in parallel with different numbers of glucose processes – naïve)
- for different periods of time (computations of different length)

Under the assumption on the environment:

- enough nutrients (water, sugar, phosphates, amino acids, nitrous bases. . .)
- no toxics
- no competing organisms (a single VICE)
- right temperature, pressure, ...

Results

Data are collected from 10^3 computations, made of 10^4 transitions, involving 10^6 different processes (~ 12 hours each)

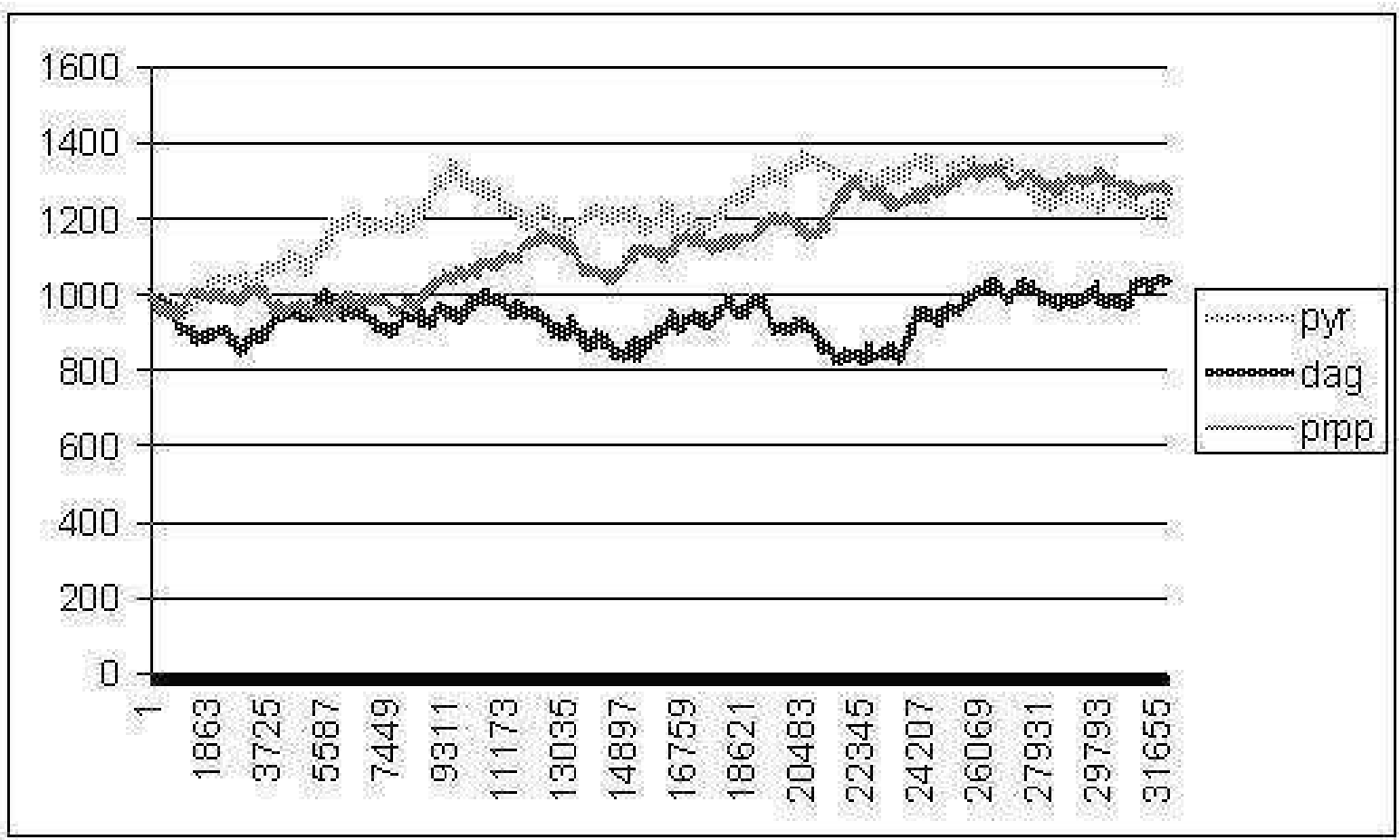
Throughput:

- Production of energy and metabolites, through oxidation of glucose, shows homeostasis
- biomass produced as expected

Distribution of metabolites over Glycolysis pathway:

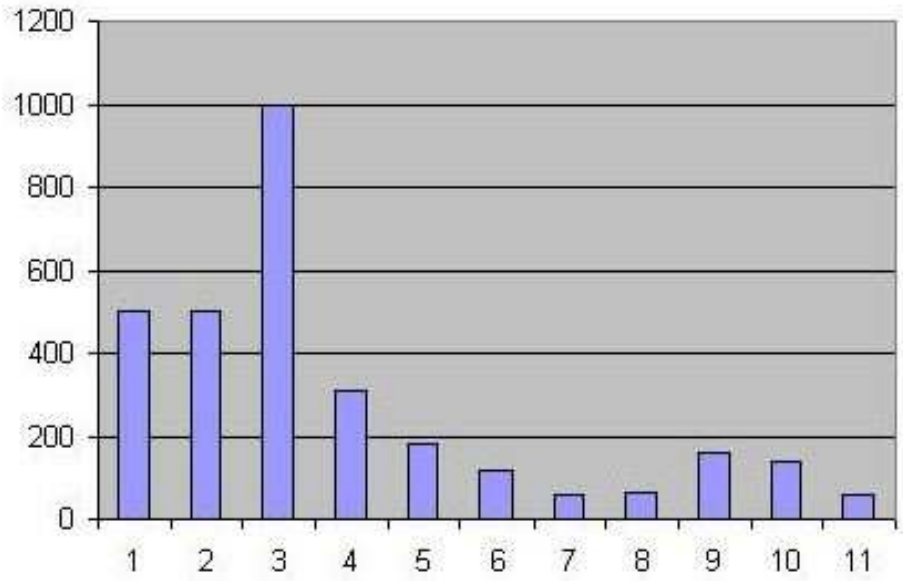
- Like in real prokaryotes (in their steady state)
- The distributions agree with those computed in vitro.

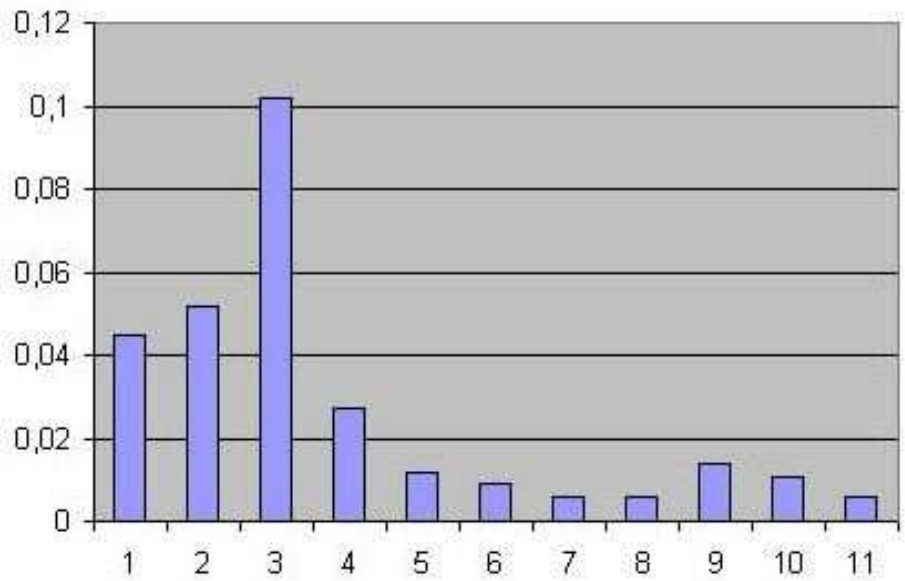
Steady state



pyruvate, diacilglycerol, phosphoribosylpyrophosphate

Usage of enzymes



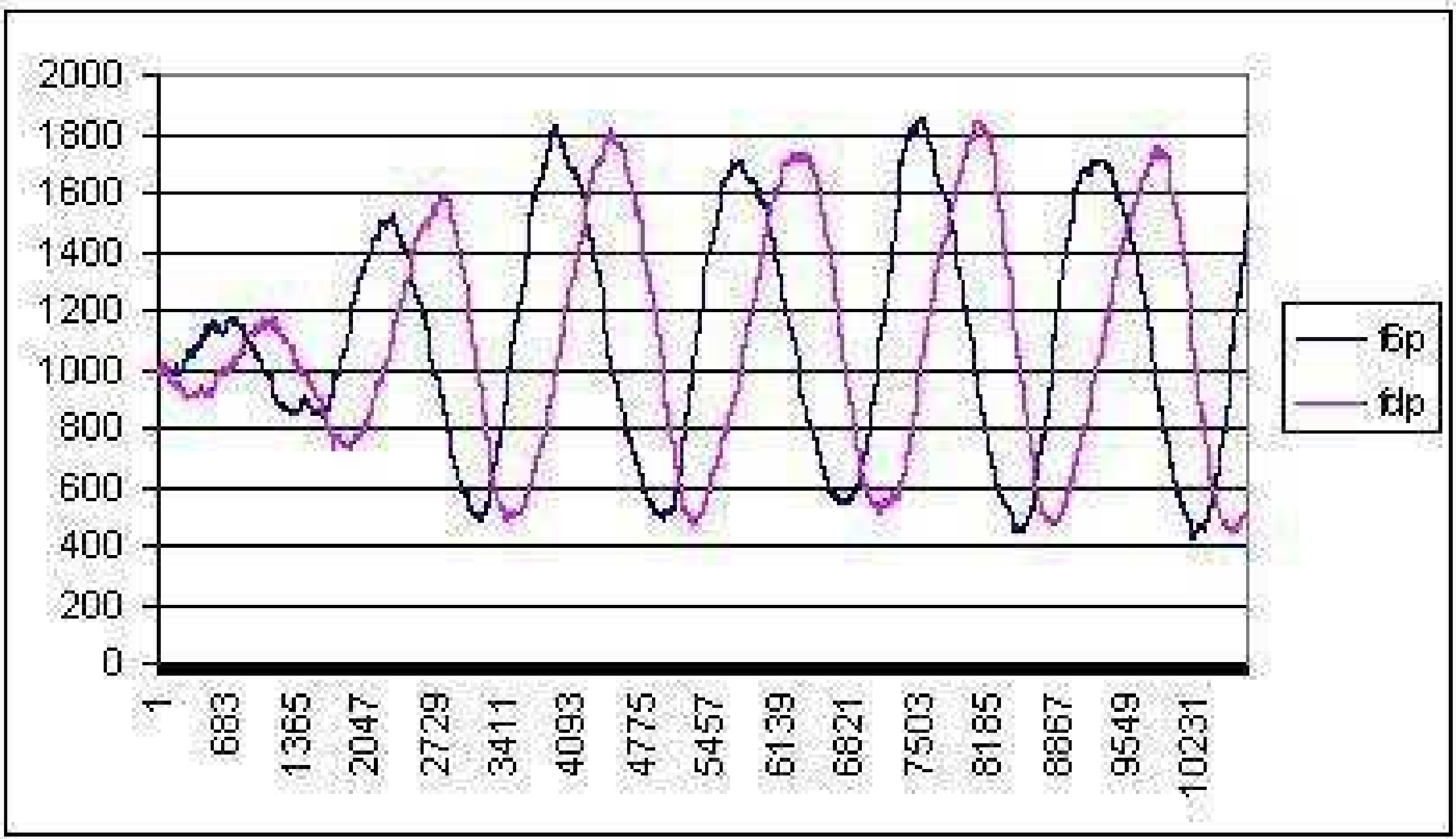


1	mg111	5	mg300	9	compl. pyr. dehydrogenase
2	mg215	6	mg430	10	mg299
3	mg023	7	mg407	11	mg357
4	mg031	8	mg216		

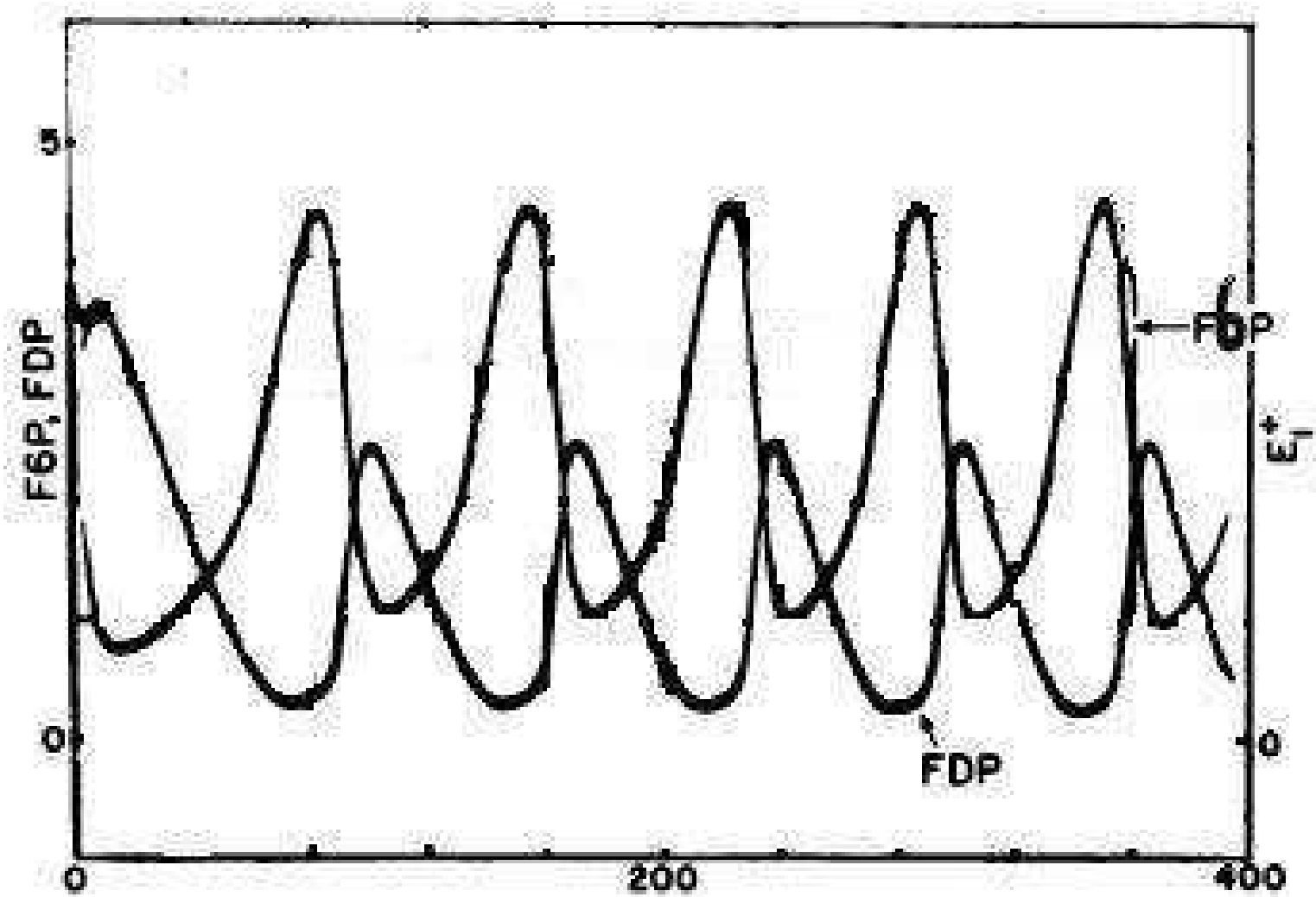
Something emerges

- Add the specification of a regulatory feedback circuit on the enzyme phosphofructokinase (the more ADP the faster the phosphorilation of fructose-6-phosphate).
  Look then at the time course of fructose-6-phosphate and fructose-1.6-bephosphate
- Change the feeding regimen by supplying the sugar:
  - all at the beginning, a huge quantity — no oscillations
  - at a constant rate — oscillations show up!!

Oscillations



the real ones ...



Other case studies ...

- Specify and run the metabolome of Escherichia coli
- Because of efficiency problems, a new implementation
  - a subset of CCS (fast also with name passing)
  - essentially multiplication of stoichiometric matrices
  - more than two orders of magnitude faster than the previous one (10^8 transitions involving 10^7 processes in less than 8 hours — done while sleeping ...)

E. Coli

- The virtual behaviour “matches” the real one
- Knock out some genes
  - agrees on known KO (ppc, pgi, zwf)
  - a new KO (rpe) – no data in the literature

Neurons

- A first step to studying plasticity and memory
- Pre-synaptic mechanisms of neuro-transmitter release
- Executable model (in Spim)
- Results agree with other deterministic, non executable models
- More and news in a few minutes during Andrea's talk

Conclusions

- Cells as processes $\Rightarrow$ "virtual" living matter
- Formal, mathematical theory $\Rightarrow$ mechanical analysis tools
  - constructive and executable
  - compositional, with different abstraction levels
- Quantities crucial for behavioural descriptions
- New computational models (e.g. new interation mechanisms) $\Rightarrow$ new semantics
- "Virtual" experiments as computations $\Leftarrow$ not enough!!

To Do

Far from satisfactory languages! New challenges:

- membranes, compartments and the like
- geometrical issues
- more faithful (and efficient) bio-chemistry
- causality
- usability (graphich interfaces, fast interpreters, specification generators from data bases, ...)
- new analysis techniques (static vs dynamic) and tools

Towards ...

Bio-calculus environment

Towards uniform (families of) environments

- sharing formal grounds and tools
- providing the user with mechanisms for describing systems at different levels of abstraction

More fundamental research and more case studies