Bioinformatics beyond sequences

Knowledge representation and analysis of biological data

Per J. Kraulis

What is bioinformatics?

- "Information technology applied to the management and analysis of biological data"
    Attwood & Parry-Smith 1999

- "Collection, archiving, organization and interpretation of biological data"
    Thornton 2003

Sequence databases

```
ID   RASH_HUMAN      STANDARD;      PRT;    189 AA.
AC   P01112; Q14080; Q6FHV9;
DT   21-JUL-1986, integrated into UniProtKB/Swiss-Prot.
DT   21-JUL-1986, sequence version 1.
DT   07-MAR-2006, entry version 77.
DE   GTPase HRas precursor (Transforming protein p21) (p21ras) (H-Ras-1)
DE   (c-H-ras).
GN   Name=HRAS; Synonyms=HRAS1;
OS   Homo sapiens (Human).

CC   -!- FUNCTION: Ras proteins bind GDP/GTP and possess intrinsic GTPase
CC       activity.
CC   -!- ENZYME REGULATION: Alternate between an inactive form bound to GDP
CC       and an active form bound to GTP. Activated by a guanine
CC       nucleotide-exchange factor (GEF) and inactivated by a GTPase-
CC       activating protein (GAP).


SQ   SEQUENCE   189 AA;  21298 MW;  EE6DC2D933E2856A CRC64;
     MTEYKLVVVG AGGVGKSALT IQLIQNHFVD EYDPTIEDSY RKQVVIDGET CLLDILDTAG
     QEEYSAMRDQ YMRTGEGFLC VFAINNTKSF EDIHQYREQI KRVKDSDDVP MVLVGNKCDL
     AARTVESRQA QDLARSYGIP YIETSAKTRQ GVEDAFYTLV REIRQHKLRK LNPPDESGPG
     CMSCKCVLS
//
```

Sequence analysis

```
17 UNIPROT:Q503B6_BRARE 1:189 1:189    REIRQHKLRKLNPPDDNGQDCMNCRCVVS
18 UNIPROT:Q568K0_BRARE 1:189 1:189    REIRQHKMRKLNPPDESGQDCMSCRCVVS
19 UNIPROT:RASK_HUMAN   1:188 1:188    REIRQYRLKKISK-EEKTPGCVKIkcII-
20 UNIPROT:Q3UCX0_MOUSE 1:188 1:188    REIRQYRMRKLNSSDDGTQGCMGLPCVL-
21 UNIPROT:RASN_MOUSE   1:188 1:188    REIRQYRLKKLNSSDDGTQGCMGSPCVL-
22 UNIPROT:RASK_MOUSE   1:188 1:188    REIRQYRLKKISK-EEKTPGCVKIkcVI-
23 UNIPROT:RASK_RAT     1:188 1:188    REIRQYRLKKISK-EEKTPGCVKIkcVI-
24 UNIPROT:Q4FJP3_MOUSE 1:188 1:188    REIRQYRMKKLNSSDDGTQGCMGLPCVL-
25 UNIPROT:Q9D091_MOUSE 1:188 1:188    REIRQYRMKKLNSSDDGTQGCMGLPCVL-
26 UNIPROT:RASN_CHICK   1:188 1:188    REIRQYRMKKLNSNEDGNQGCMGLSCIV-
27 UNIPROT:RASN_HUMAN   1:188 1:188    REIRQYRMKKLNSSDDGTQGCMGLPCVV-
28 UNIPROT:Q5U091_HUMAN 1:188 1:188    REIRQYRMKKLNSSDDGTQGCMGLPCVV-
29 UNIPROT:Q2MJK3_PIG   1:188 1:188    REIRQYRMKKLNSSDDGTQGCMGLPCVV-
30 UNIPROT:RASN_CAVPO   1:188 1:188    REIRQYRMKKLNSNDDGTQGCMGLPCVV-
31 UNIPROT:Q4S7E9_TETNG 1:188 1:188    REIRQYRLNKLSK-EEKTPRCVKlkCVV-
32 UNIPROT:Q3TMF4_MOUSE 1:188 1:188    REIRQYRMKKLNSSDDGTQGCMGLPCVL-
33 UNIPROT:RASN_RAT     1:188 1:188    REIRQYRMKKLNSSEDGTQGCMGLPCVV-
34 UNIPROT:RASN_MONDO   1:188 1:188    REIRQYRMKKLNSSDDGTQGCLGLSCAV-
35 UNIPROT:RASN_PONPY   1:188 1:188    REIRQYRMKKLNSSDDGTQGCMGLPCVV-
36 UNIPROT:O57467_ORYLA 1:188 1:188    REIRQYRLSKLSK-EEKTPRCVNLkcVV-
37 UNIPROT:O13021_XENLA 1:185 1:184    REIRQFRLKKMSK-EEKTPGCVKFK----
38 UNIPROT:Q5EFX7-2     1:188 1:188    REIRQYRLSKISK-EEKTPGCVQLkcVV-
39 UNIPROT:RASN_XENLA   1:188 1:188    REIHQYRMKKLDSSEDNNQGCIRIPCKL-
40 UNIPROT:RASK_MSVKI   1:188 1:188    REIRQYRLKKISK-EEKTPGCVKIkcVI-
41 UNIPROT:RAS_CARAU    1:177 1:177    REIRQYRLRKLSKEEET------------
42 UNIPROT:Q6DGD1_BRARE 1:186 1:185    REIRHYRMKKLNSREDRKQGCLGVSC---
43 UNIPROT:P01116-2     1:188 1:187    REIRKHK-EKMSKDGKKKKKKSKTKCVI-
44 UNIPROT:RASK_MELGA   1:188 1:187    REIRKHK-EKMSKDGKKKKKKTKTKCII-
45 UNIPROT:RASK_CYPCA   1:188 1:187    REIRKHK-EKMSKEGKKKKKKSKTKCVL-
46 UNIPROT:RASK_ORYLA   1:188 1:187    REIRKHK-EKMSKEGKKKKKKSKTKCIL-
47 UNIPROT:Q9PSS8_PLAFE 1:188 1:187    REIRKHK-EKMSKEGKKKKKKSKTKCSL-
48 UNIPROT:RASK_MONDO   1:188 1:187    REIRKHK-EKMSKDGKKKKKKSKTKCII-
49 UNIPROT:RASN_BRARE   1:186 1:185    REIRHYRMKKLNSREDRKQGCLGVSC---
50 UNIPROT:Q6AZA4_BRARE 1:188 1:187    REIRKHK-EKMSKEGKKKKKKSKTKCAL-
   consensus/100%                     REI+pa+.pKhs..tct............
   consensus/90%                      REIRpa+.cKhs..tctt.tp.th.Chl.
   consensus/80%                      REIRQa+h+Kls.--tt.sChth.Cll.
   consensus/70%                      REIRQa+h+KLs.s--ps.GChthpCVl.
```
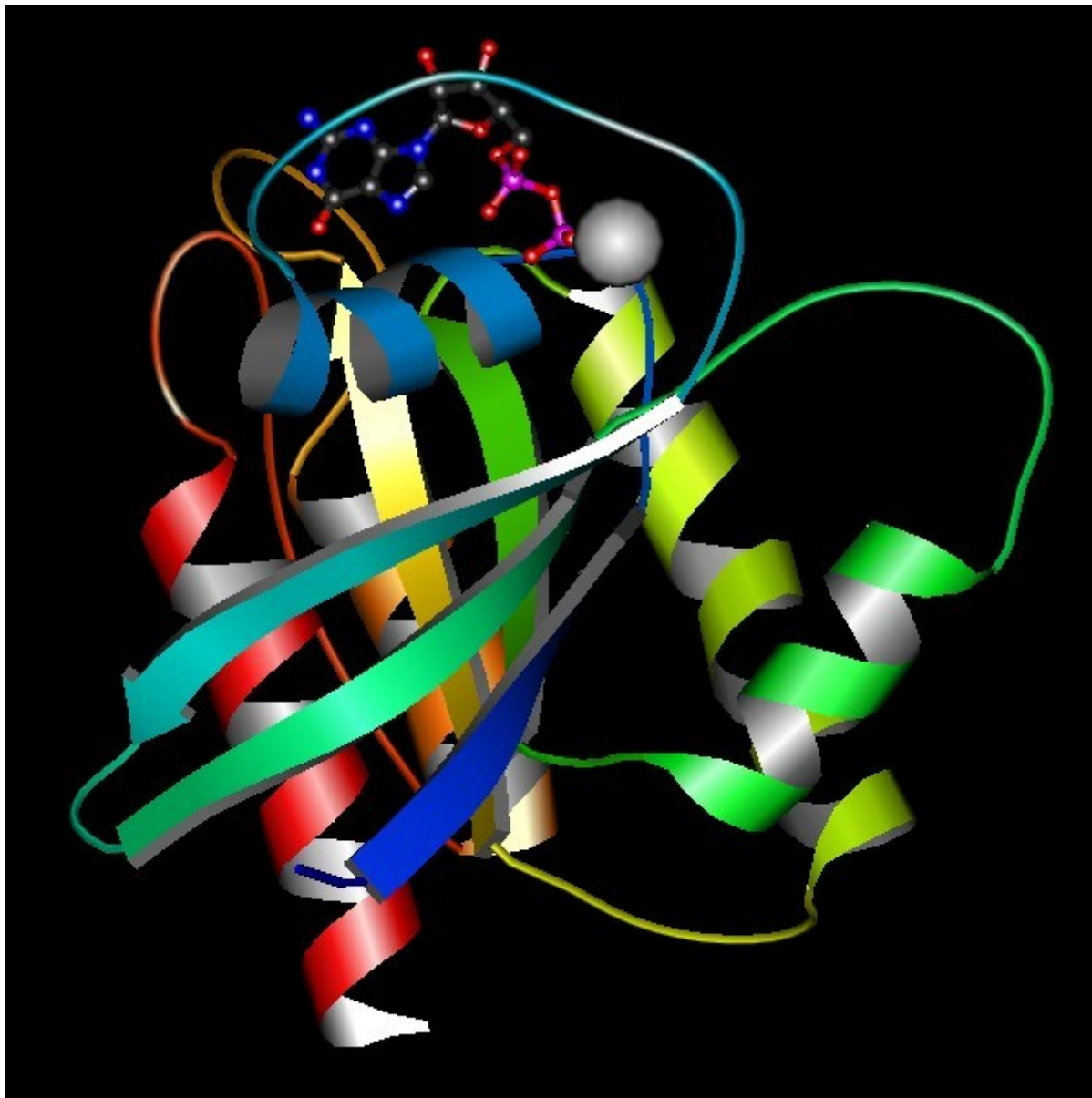
MolScript: Per Kraulis 1991, 1997

METHIONINE METABOLISM

KEGG: Kanehisa 2004
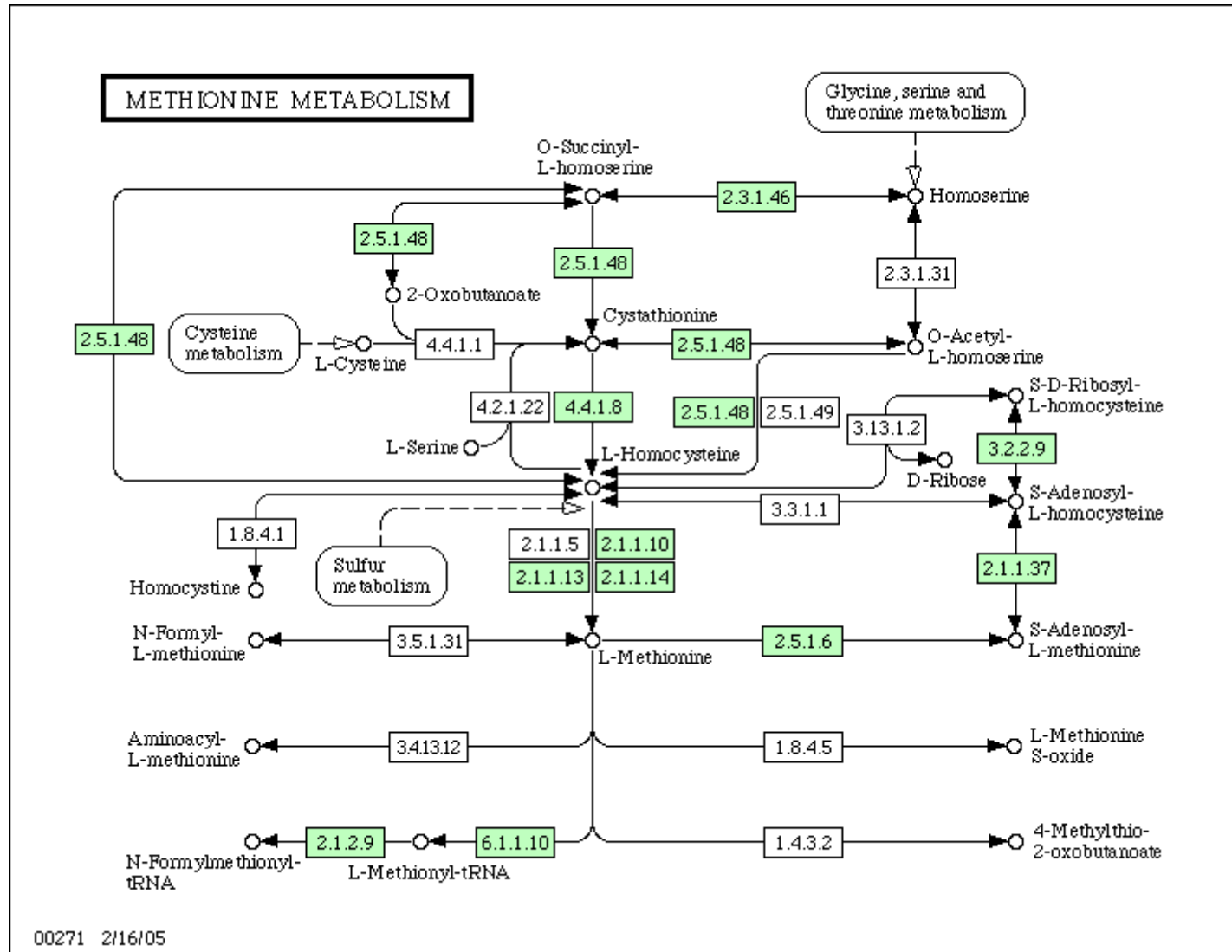
Knowledge Representation (KR)

- Biomedicine: "Difficult" data
  - Different scales (molecules … organisms)
  - Complexity: objects, relations

- Usage should govern representation
  - Searching: find relevant info
  - Analysis: e.g. comparison
  - Computation: simulation

Project 1:
Improved data model for pathways

- Molecular states
- Complexes
- Locations
- Events
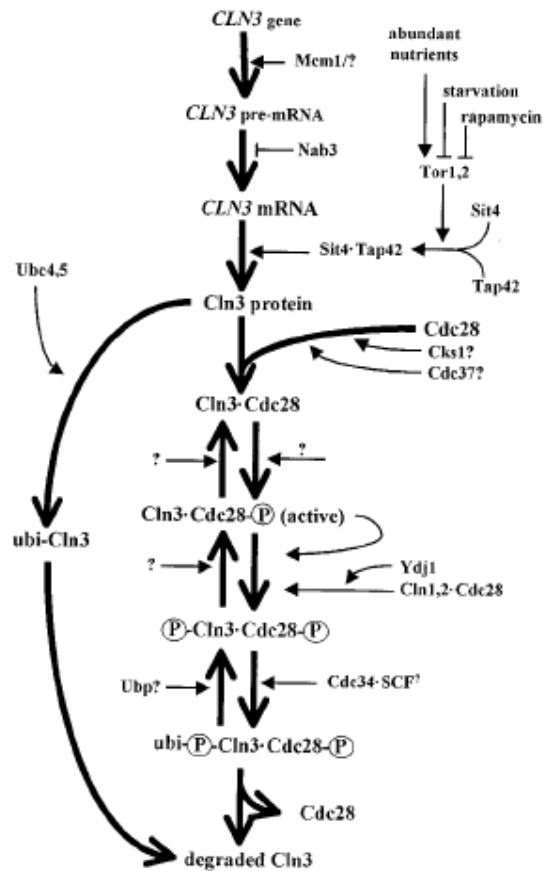- Hierarchy; levels of detail
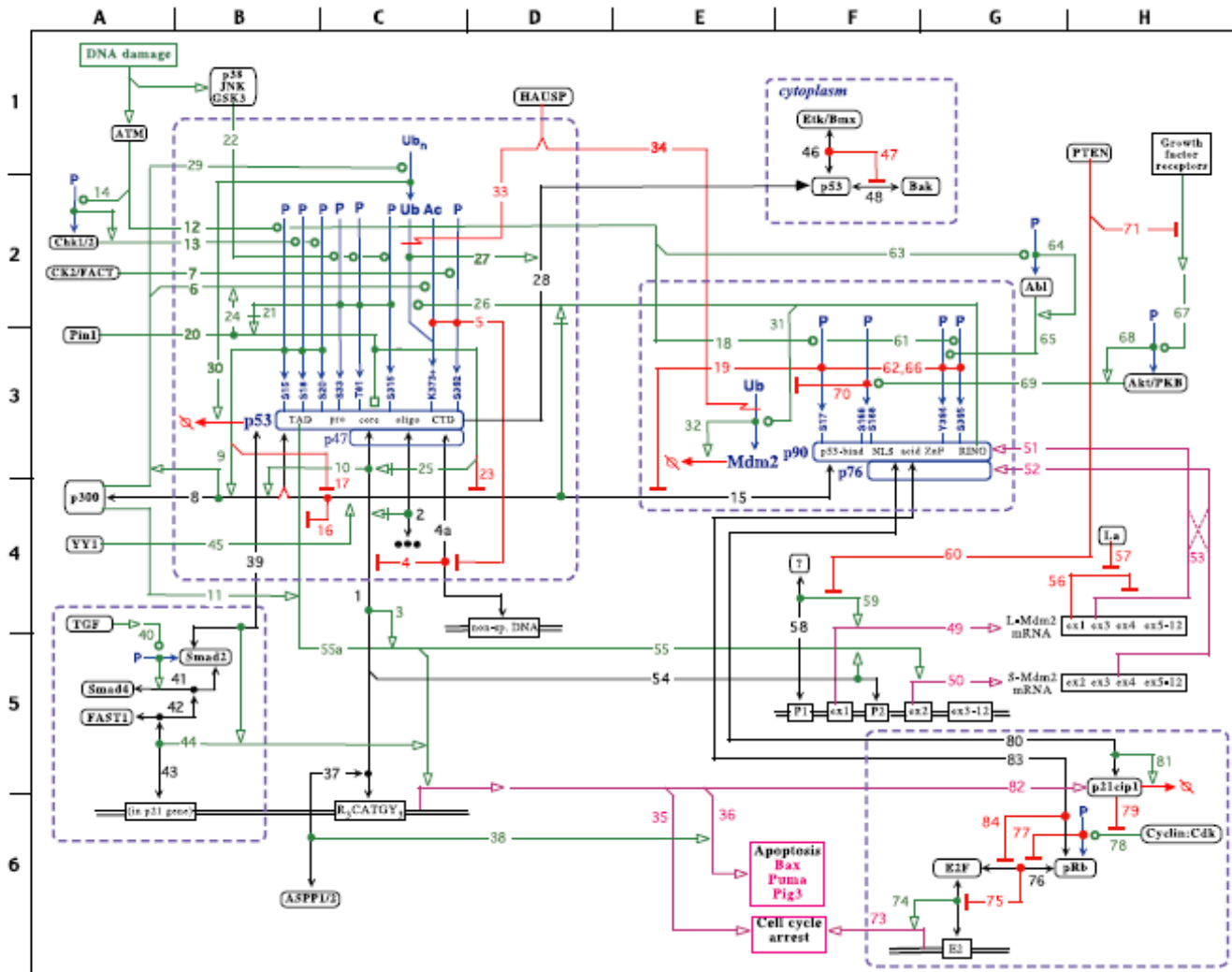
p53 and Mdm2 interactions: Kohn & Pommier 2005

Statecharts

- David Harel, 1987
- State-transition diagrams, extended with
  - Hierarchy
  - Orthogonality
  - Communication
- For reactive systems
  - Event-driven
  - Stimuli; external and internal

GeneCV

Mendenhall & Hodge 1998

- The life of a biomolecule
- Objects
  - Gene
  - Protein
  - Complexes
  - Locations
- Events
  - Creation
  - Destruction
  - Regulation
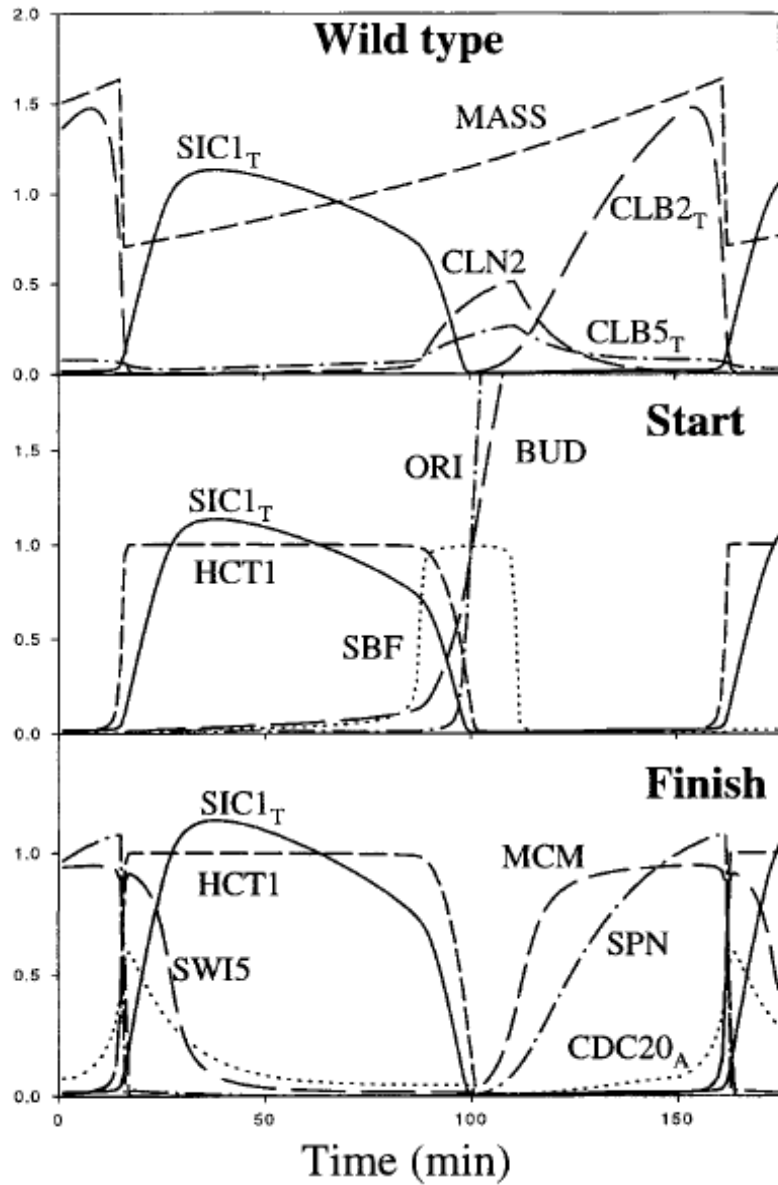 - Transport
  - Interaction
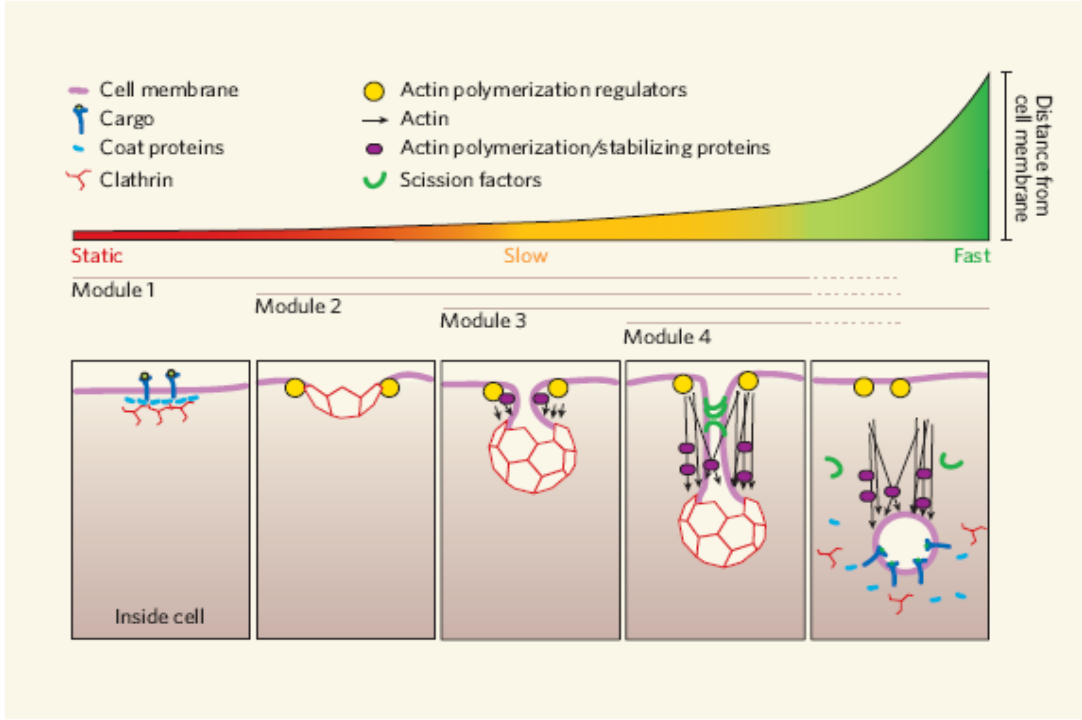- Statecharts

Project 2:
Data model for biological processes

- Temporal data
- Events
- Activities
- Trajectories of parameters (levels)
- Temporal relationships (before, after…)
- General; allow different scales
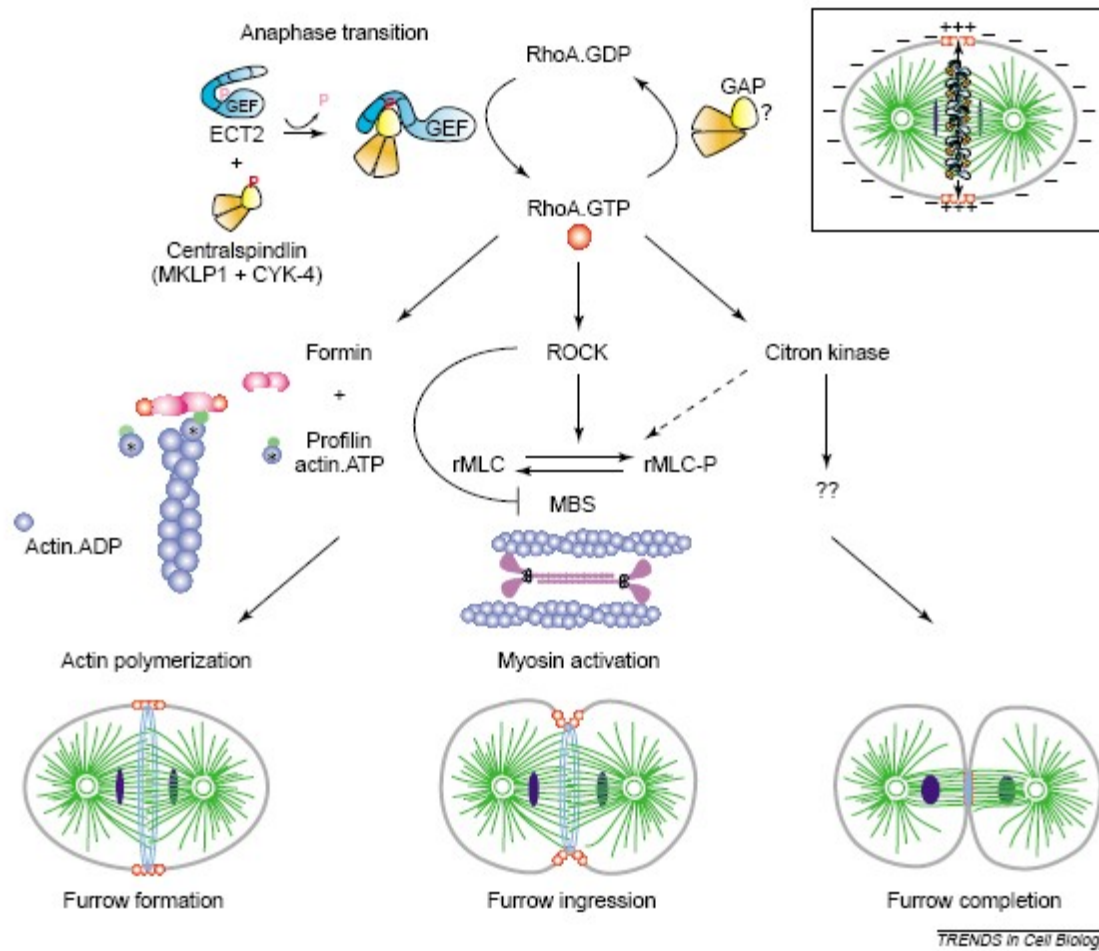
Cytokinesis: Rho regulation
Piekny, Werner, Glotzer 2005

Kinetic analysis of budding yeast cell cycle: Chen et al 2000

Endocytic vesicle formation
Duncan & Payne 2005

The Chronicle system

- Temporal database
- Macroscopic systems
  - Cells
  - Signaling cascades
  - *In vivo* studies
- Inspired by Geographical Information Systems (GIS) research
- Prototype: Sara Eriksson, Biovitrum