GeneCV

Knowledge Representation of Molecular Events in Cellular Pathways

Per Kraulis

Biological processes

- Examples:
  - – Pathways
  - – Signaling
  - – Cell cycle
  - – Growth of an organism
  - – Response of a system to perturbations
- Currently, no standard DB design
- Several new ideas, initiatives…

Biological complexity

- Biological data are inherently complex
- More complex than physics, astronomy
- Several different sources of complexity
- Data handling: difficult, serious problem

Levels of abstraction

- Molecular structures
- Polymers (DNA, protein)
- Features: genes, domains,…
- Complexes
- Cells, compartments
- Tissues, and so on…
- → Makes data modelling hard

No law without an exception

- There are very few biological natural laws
- Most proper laws from physics, chemistry
- Many common mechanisms and structures
- But, always an exception somewhere
- → Makes data modelling hard

Knowledge Representation

- Ontology
  - What exists: entities, objects, items
  - What relationships: associations, relations
- Important tools:
  - Is-a relationships: classes, inheritance
    - JNK1 is a kinase, is an enzyme, is a protein
  - Part-of relationship: composition
    - Nucleosome: DNA + histone octamer

Criteria for Knowledge Representation systems

- Match the scientist’s view of the universe
  - Use domain-specific terms, concepts
  - Avoid novel or alien concepts
- Focused: clear domain definition
- Formalize information
  - Allow computation
  - Allow database, publication
  - Implement uncertainty, updates, deletions

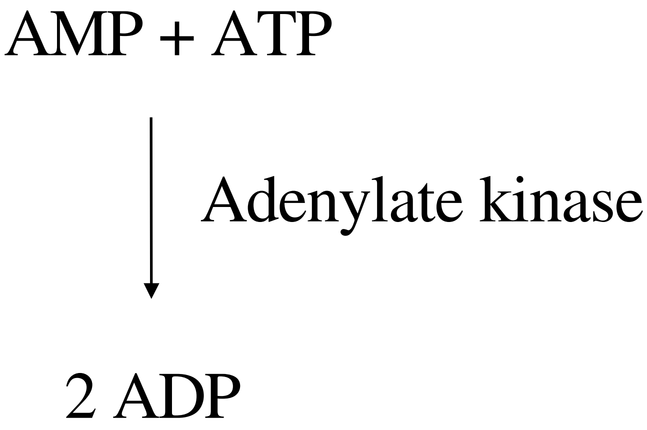
What data is represented, and how

- Explicit data model should be required
- Self-evident? No.
  - Many DBs have unclear semantics
  - Implicit assumptions are dangerous
  - Important additional data overlooked

Choice of data model has consequences

- Directly
  - Missing data
    - Some compounds are considered implicit
  - Conflated entities
    - Is ”Fe” in KEGG Fe^{2+}or Fe^{3+} ?
- Indirectly
  - Some analysis becomes harder
  - Constrains future extensions

Metabolic pathway DBs



AMP + ATP

Adenylate kinase

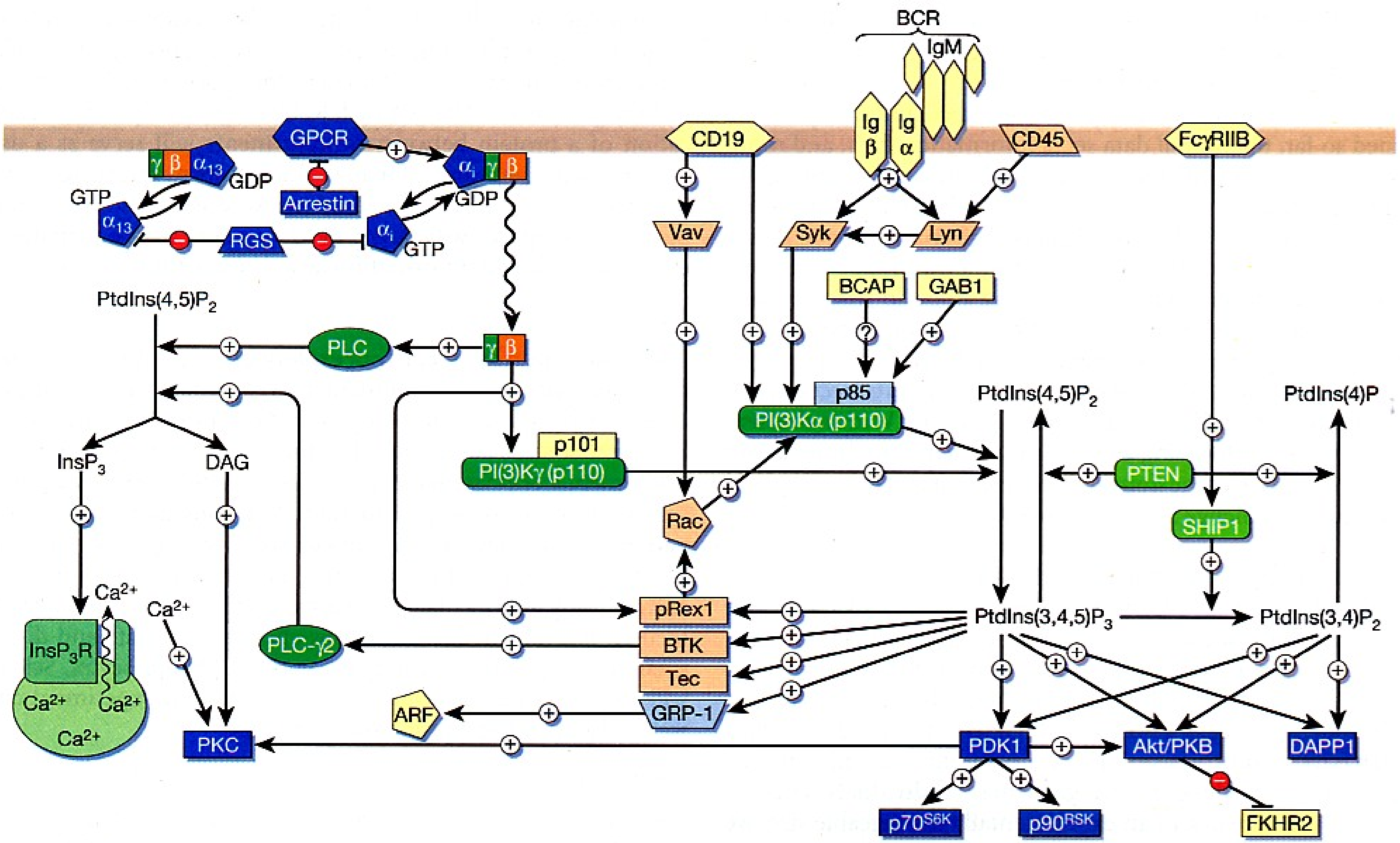
2 ADP

- Missing data (typically):
  - Species
  - Cellular location
  - During what processes?
  - Kinetic parameters
  - Literature refs

Metabolic pathway DBs: problems

- More chemical than biological
  - Enzymes, proteins do not have states
  - Ill-defined connection to life processes
- No notion of classes or inheritance
  - Compounds, enzymes, reactions: that’s it
- Weak description of relationships
  - Homology?
  - Arbitrary pathway demarcations

Signaling pathway description

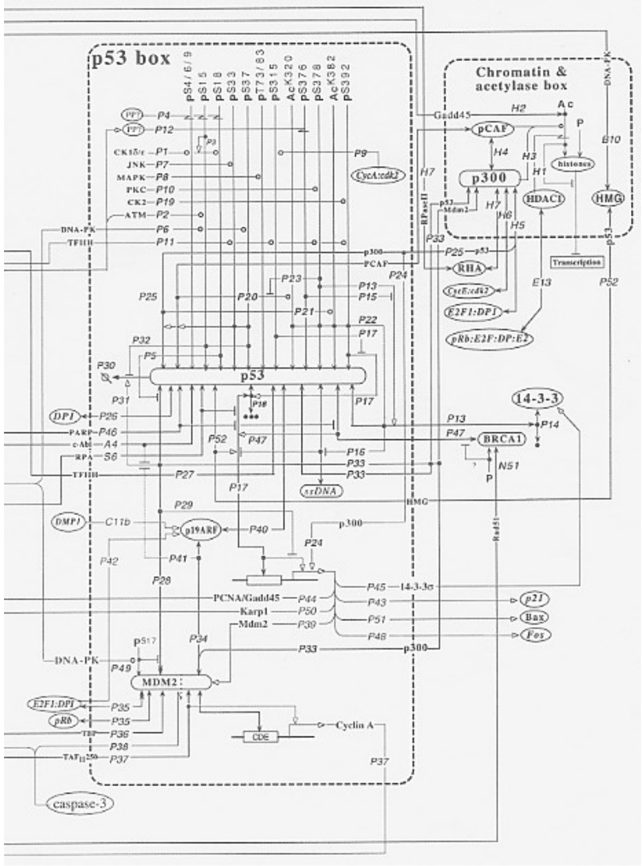


Sambrano et al, Nature (2002) 420, 708

Proteomics DBs

- Interactions between proteins
  - Literature; review-like (BIND)
- Oriented towards omics data
  - Experimental values (DIP)
- Weak connection to metabolic DBs
- Proteins only, usually

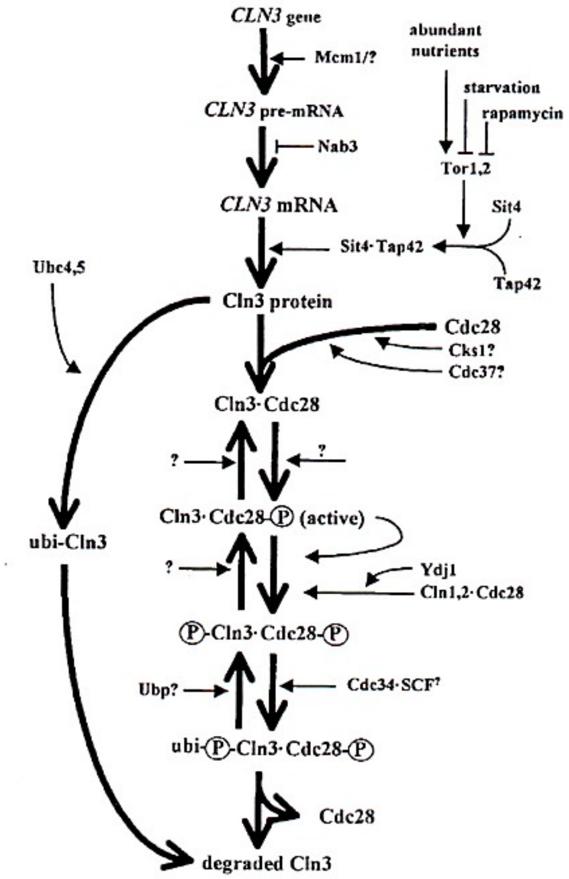
Kohn interaction maps



- Kohn, Mol Biol Cell (1999) 10, 2703-2734
- Representation of networks
- Proteins, complexes, modifications
- No dynamics
- Map only, no DB
- Failure, but interesting

What's missing?

- Proteins are changed during processes: states
- Not only proteins: RNA, molecules, complexes
- Why separate signaling and metabolism?
- The life and times of a gene, protein, molecule…



Mendenhall & Hodge, Microbiol Mol Biol Rev (1997) 62, 1191-1243

GeneCV

- Model genes, proteins, complexes
- Follow the life of a gene product
- Molecular events
- Refer to cellular process; do not model explicitly (for now)
- Based on Statecharts
- Consider: class relationships

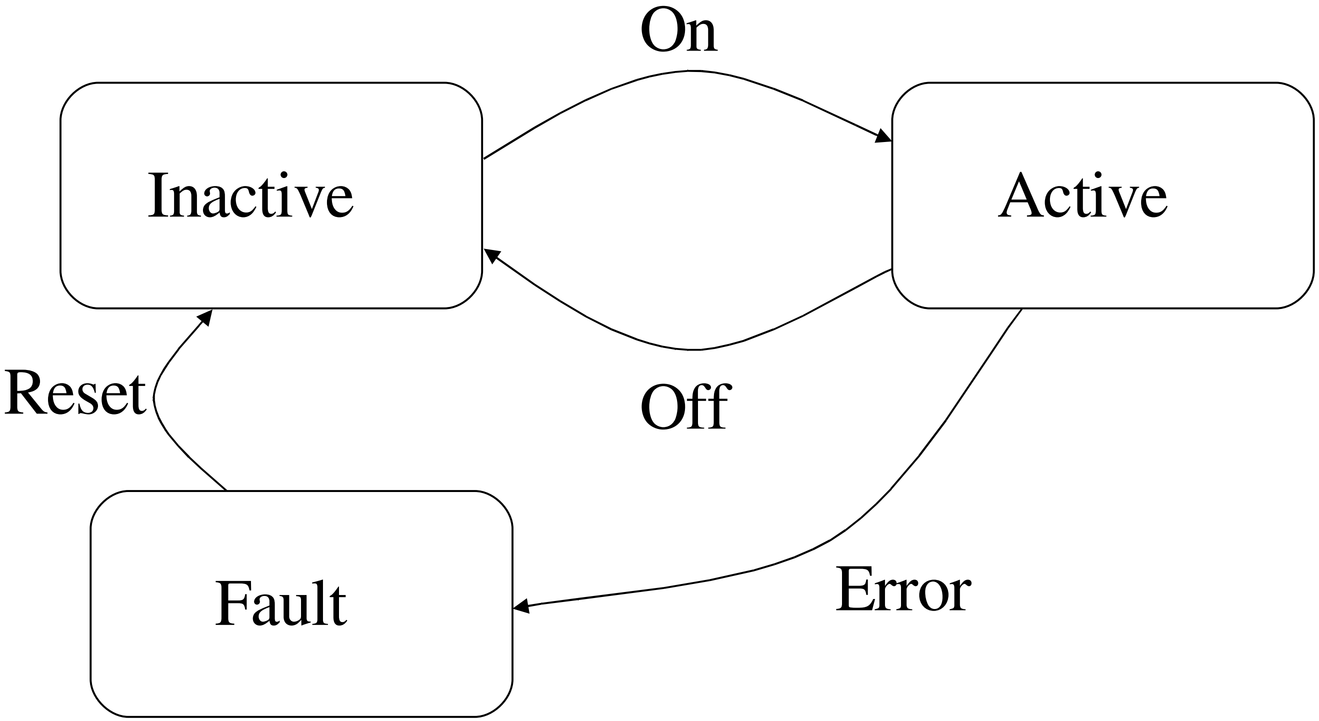
Statecharts

- David Harel 1987
- State-transition diagrams, extended with
  - Hierarchy
  - Orthogonality
  - Communication
- Designed for large reactive systems: event-driven, reacting to external and internal stimuli
- Now part of UML

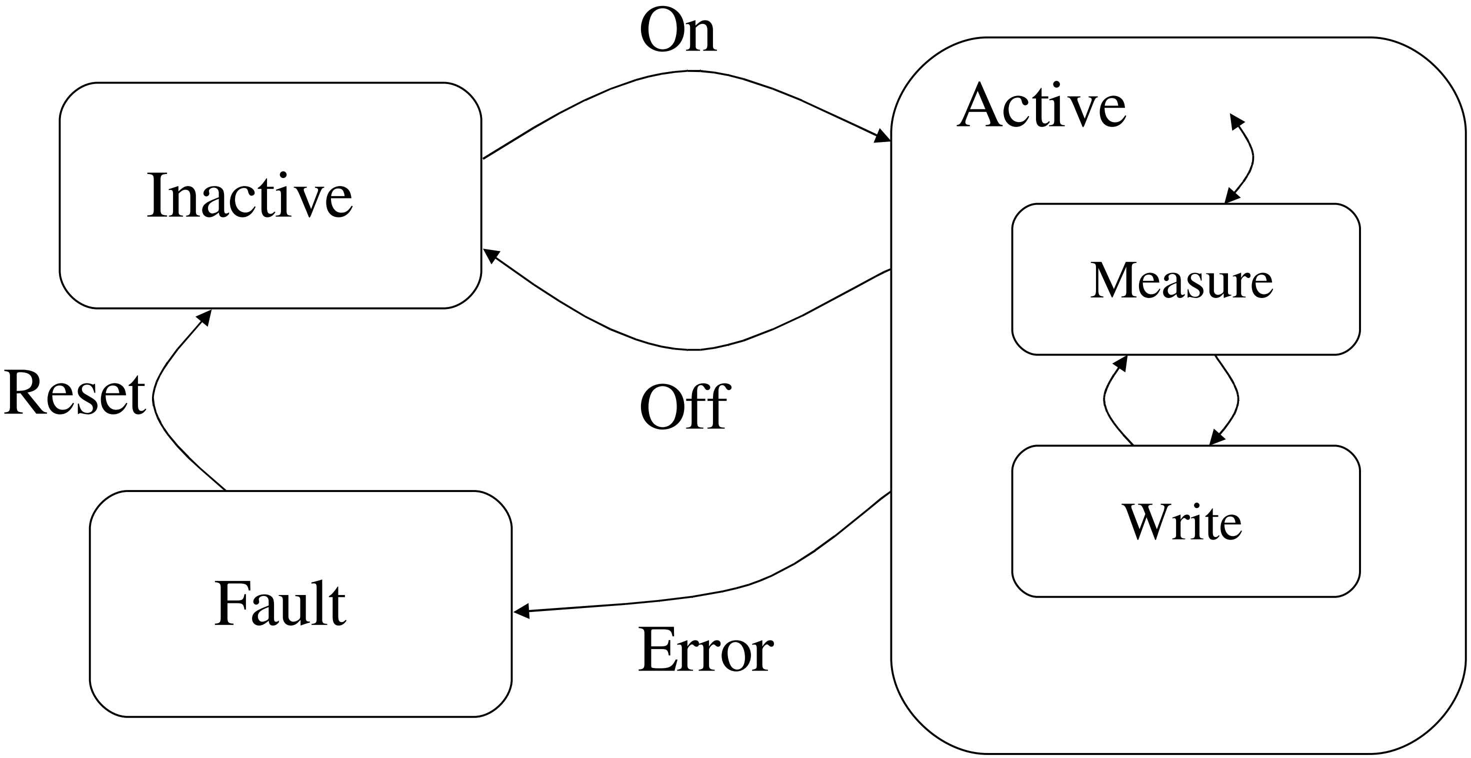
Statecharts literature

- David Harel & Michal Politi, *Modeling Reactive Systems with Statecharts*, McGraw-Hill (1998).
- David Harel, Statecharts: *A visual formalism for complex systems*, Sci Comp Prog (1987) 8, 231-274.
- Naaman Kam, David Harel, Irun R Cohen, *Modeling Biological Reactivity: Statecharts vs Boolean Logic*, Proc 2nd Conf Systems Biology (Nov 2001) Pasadena, CA, USA

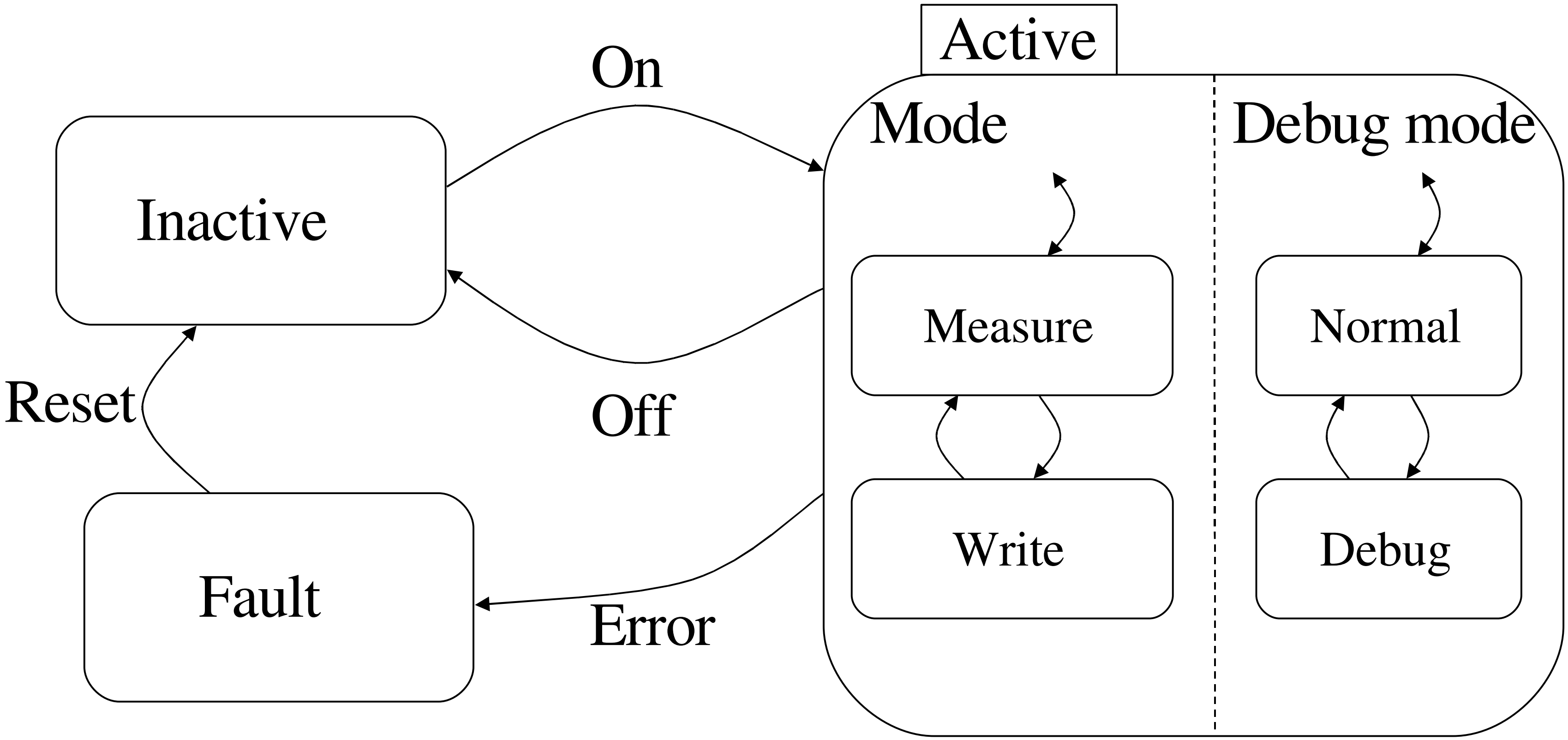
Statecharts: states and events



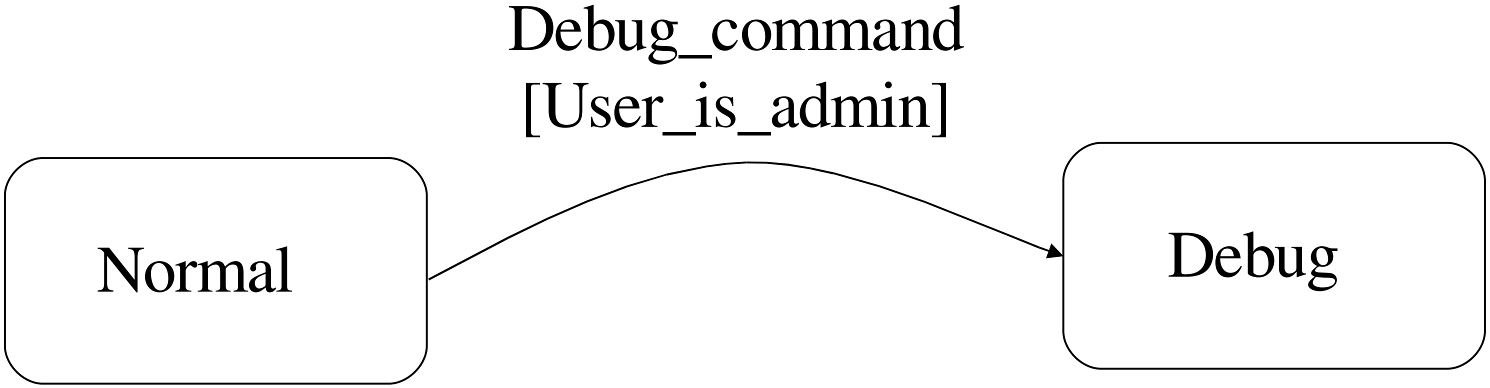
Statecharts: state hierarchy



Statecharts: state orthogonality



Statecharts: conditions

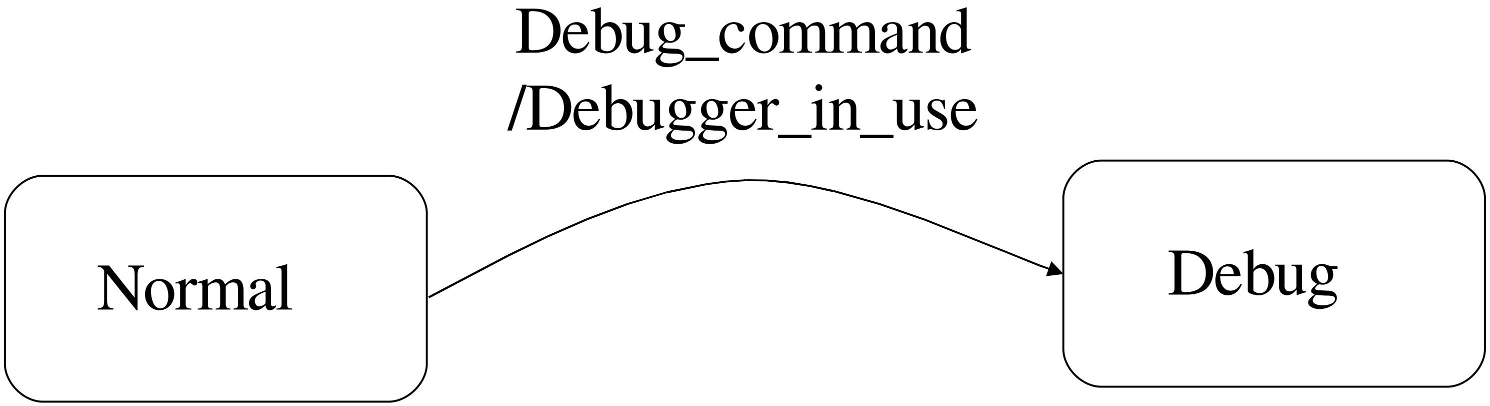


Debug_command [User_is_admin]

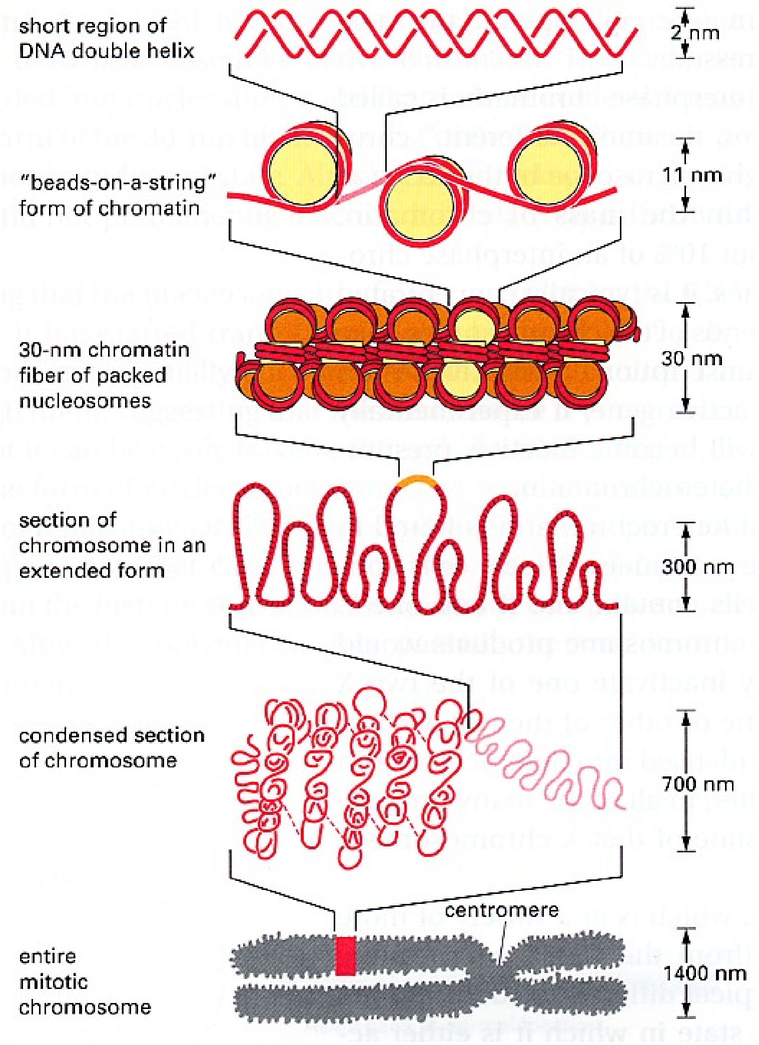
Normal

Debug

Statecharts: actions

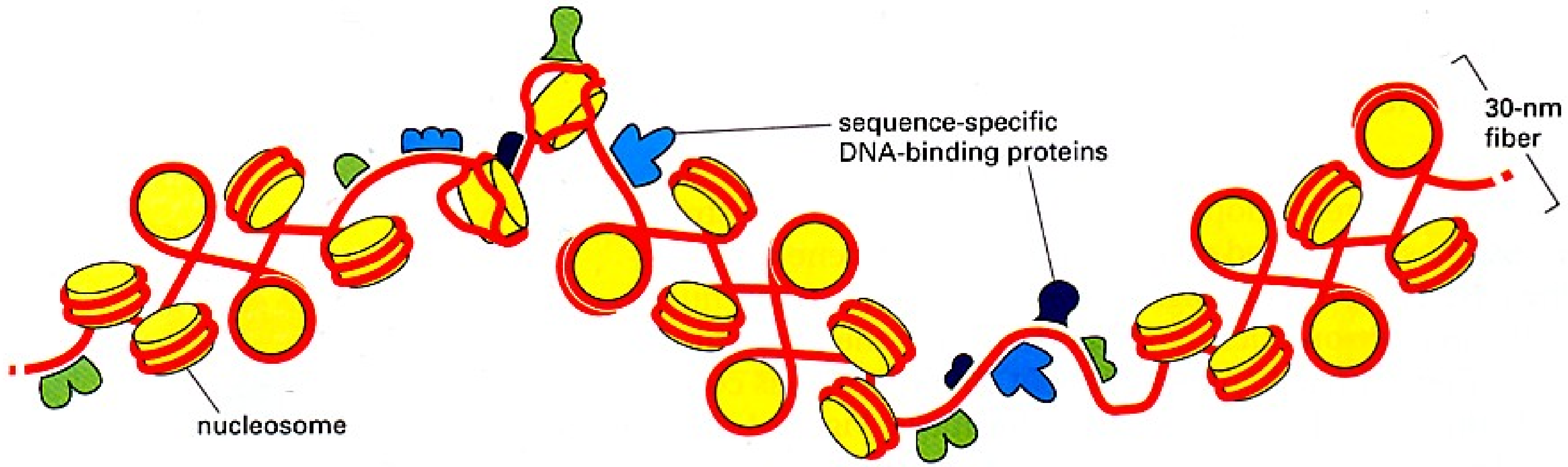


Nucleosomes: from DNA to chromosome


short region of DNA double helix
2 nm
"beads-on-a-string" form of chromatin
11 nm
30-nm chromatin fiber of packed nucleosomes
30 nm
section of chromosome in an extended form
300 nm
condensed section of chromosome
700 nm
centromere
entire mitotic chromosome
1400 nm

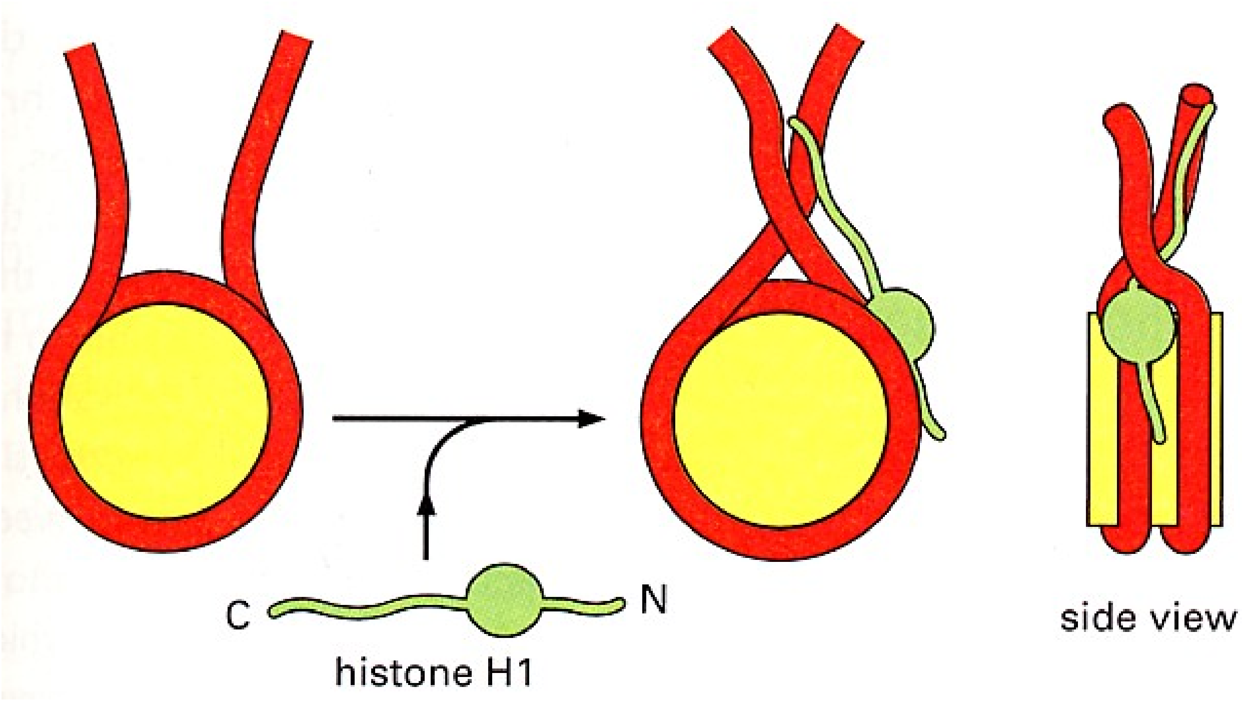
Alberts et al, Essential Cell Biology (1998) Garland

Dynamic nucleosomes



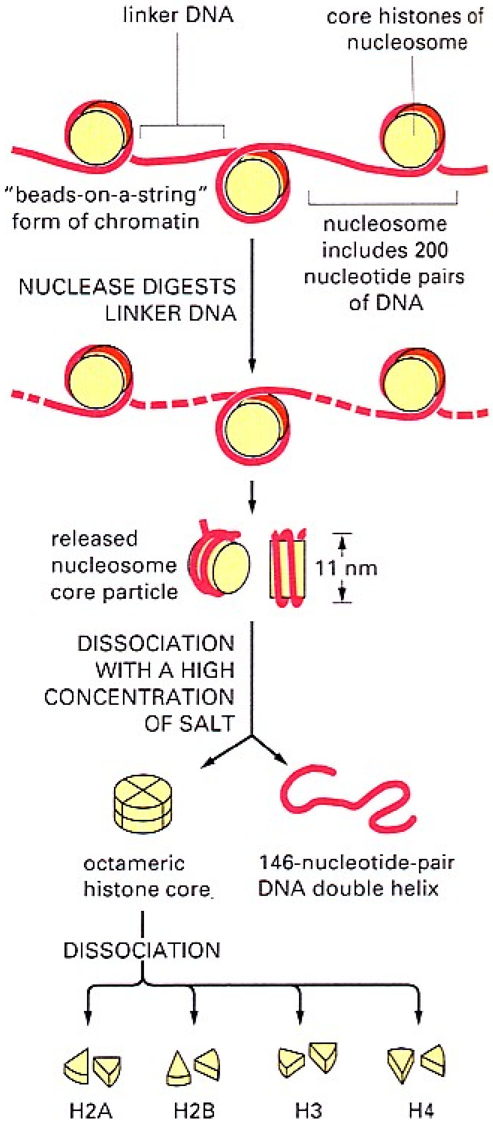
Alberts et al, Molecular Biology of the Cell (2002) Garland

Nucleosome structure



Alberts et al, Molecular Biology of the Cell (2002) Garland

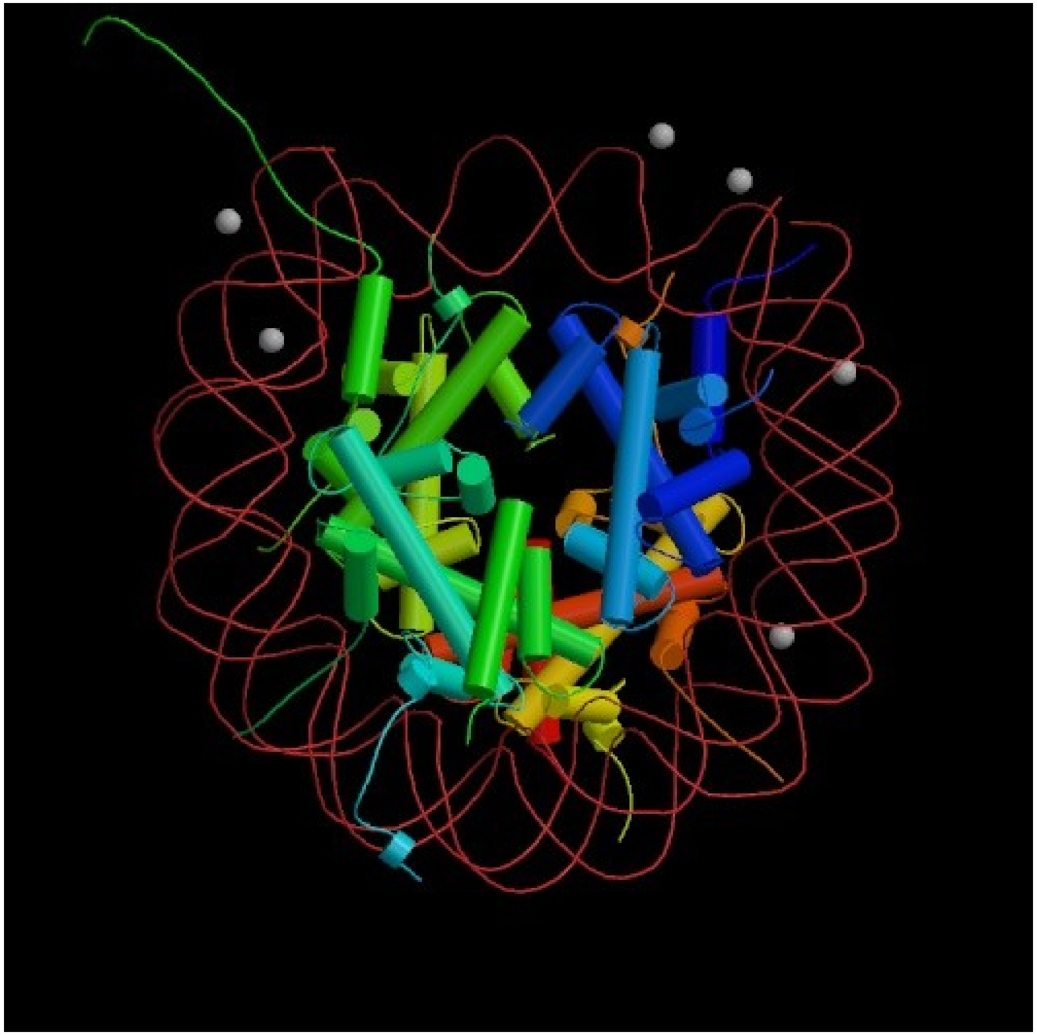
Nucleosome components



linker DNA
core histones of nucleosome
"beads-on-a-string" form of chromatin
nucleosome includes 200 nucleotide pairs of DNA
NUCLEASE DIGESTS LINKER DNA
released nucleosome core particle
11 nm
DISSOCIATION WITH A HIGH CONCENTRATION OF SALT
octameric histone core
146-nucleotide-pair DNA double helix
DISSOCIATION
H2A
H2B
H3
H4

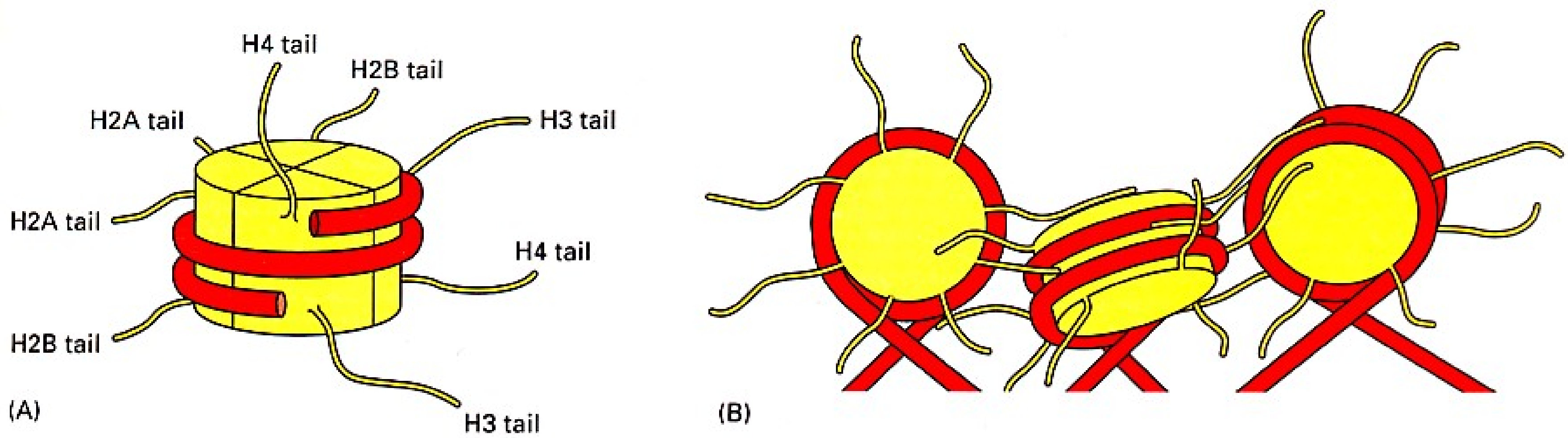
Alberts et al, Essential Cell Biology (1998) Garland

Nucleosome X-ray structure



Luger et al, Nature (1997) 389, 251

Histone tails



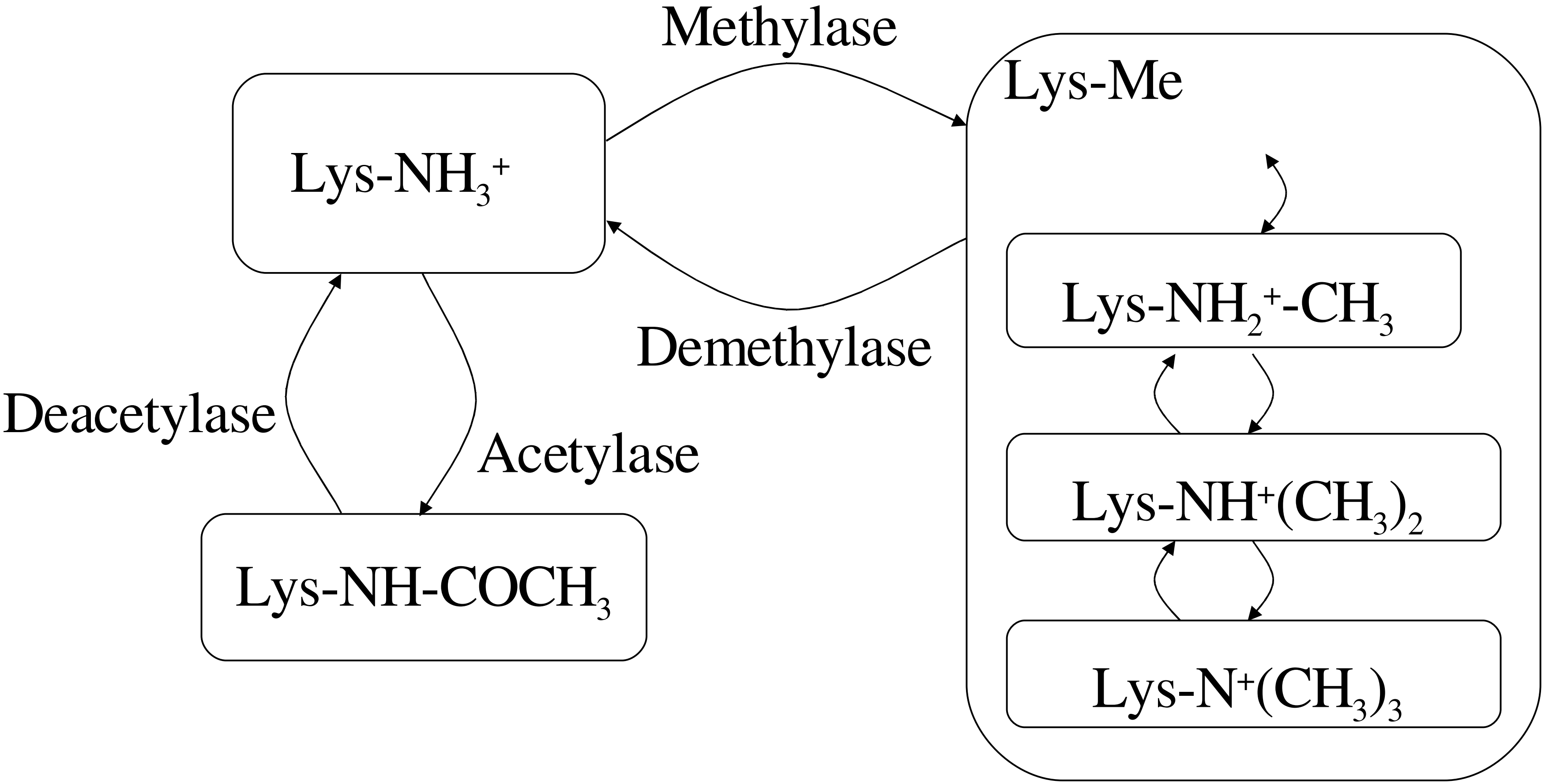
Alberts et al, Molecular Biology of the Cell (2002) Garland

Covalent modifications
= Posttranslational modifications (PTMs)

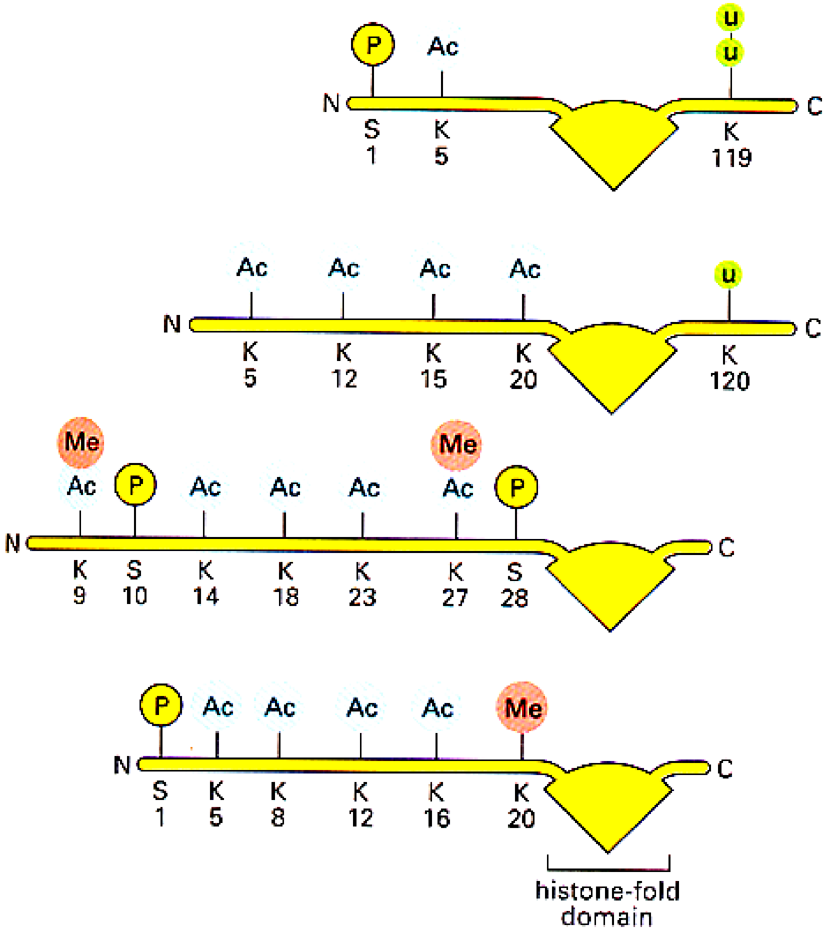
- Chemical structure is modified
- Done by enzymes
- Some reversible, others irreversible
- Binding properties change: protein complex formation

Phosphate -OPO_3	Ser, Thr Tyr His	Kinase Phosphatase
Acetyl -Ac -$COCH_3$	Lys	Acetylase Deacetylase
Methyl -Me -CH_3	Lys Arg	Methylase Demethylase
Ubiquitin -Ubq	Lys	Ubiquitin ligase

Example PTM Statechart: Lys

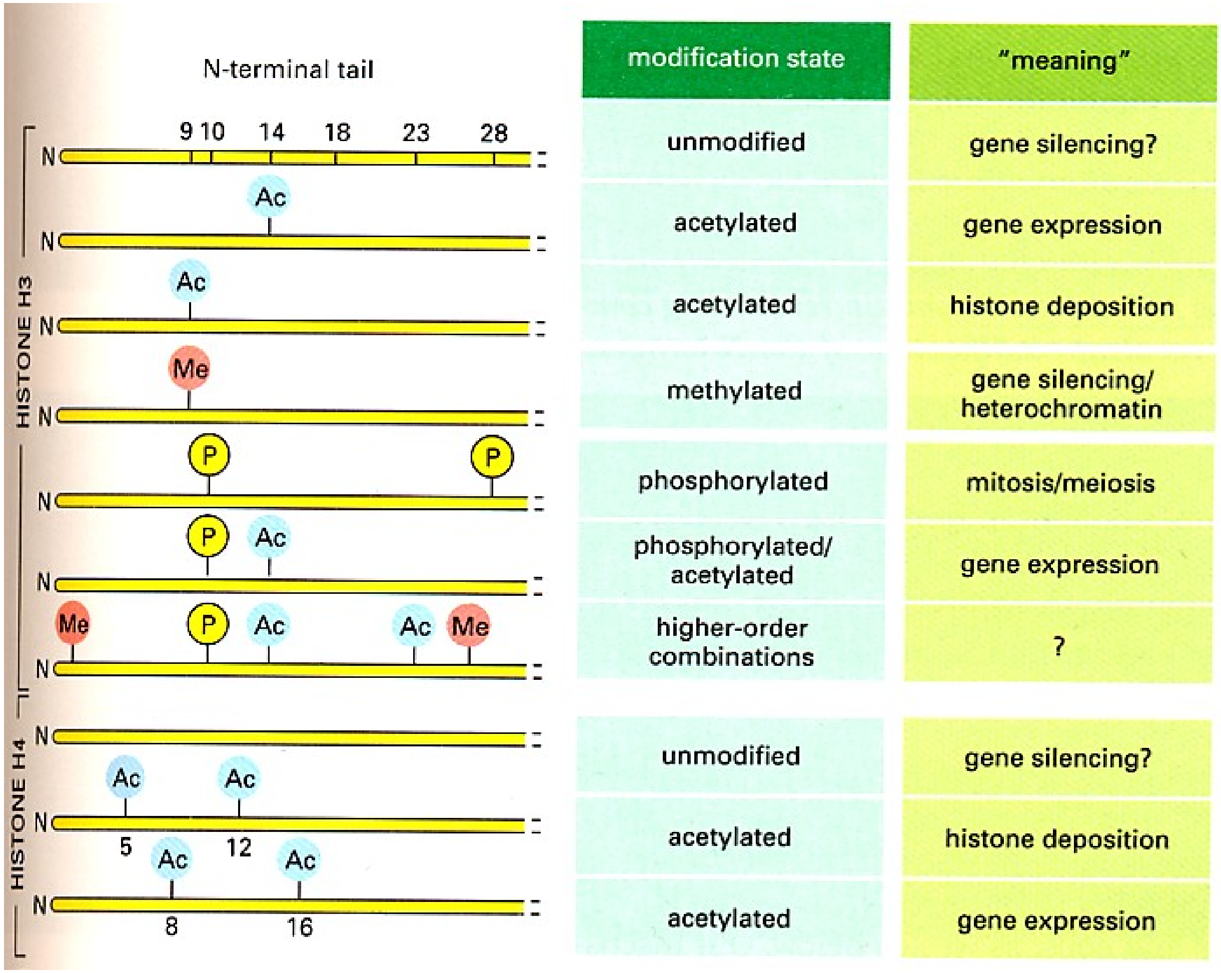


Histone tail modifications



Alberts et al, Molecular Biology of the Cell (2002) Garland

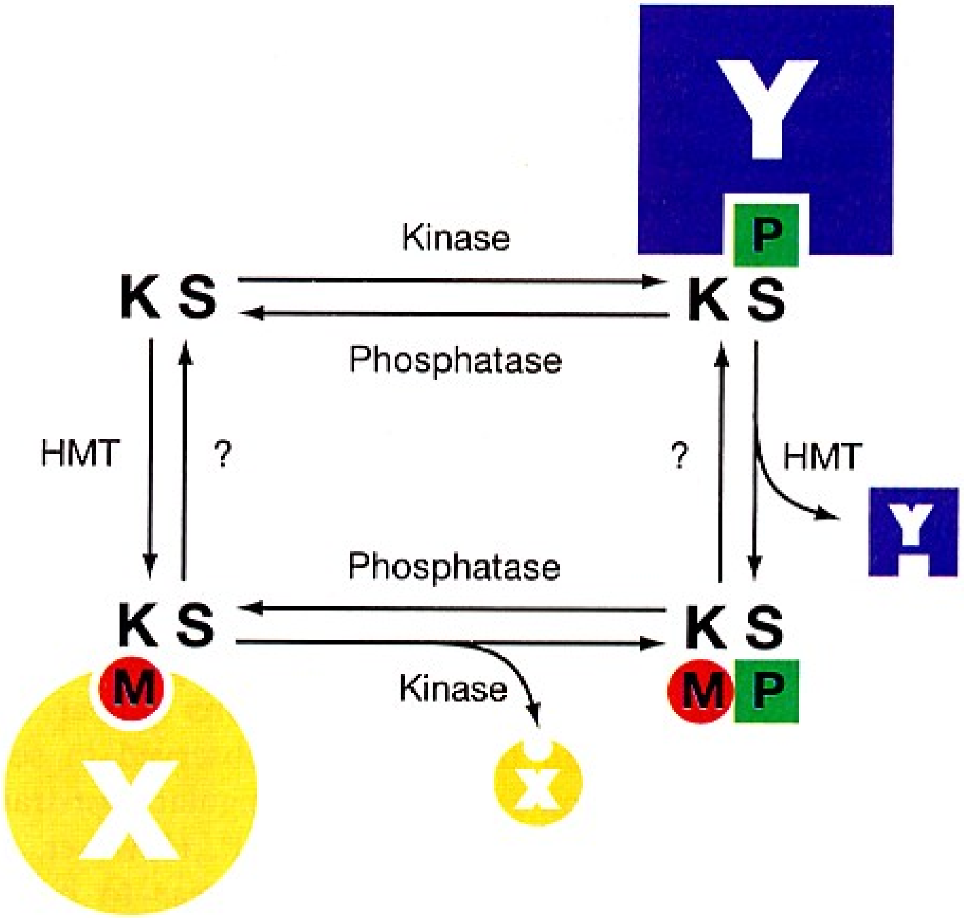
Functions of histone tail marks



modification state	"meaning"
unmodified	gene silencing?
acetylated	gene expression
acetylated	histone deposition
methylated	gene silencing/ heterochromatin
phosphorylated	mitosis/meiosis
phosphorylated/ acetylated	gene expression
higher-order combinations	?
unmodified	gene silencing?
acetylated	histone deposition
acetylated	gene expression

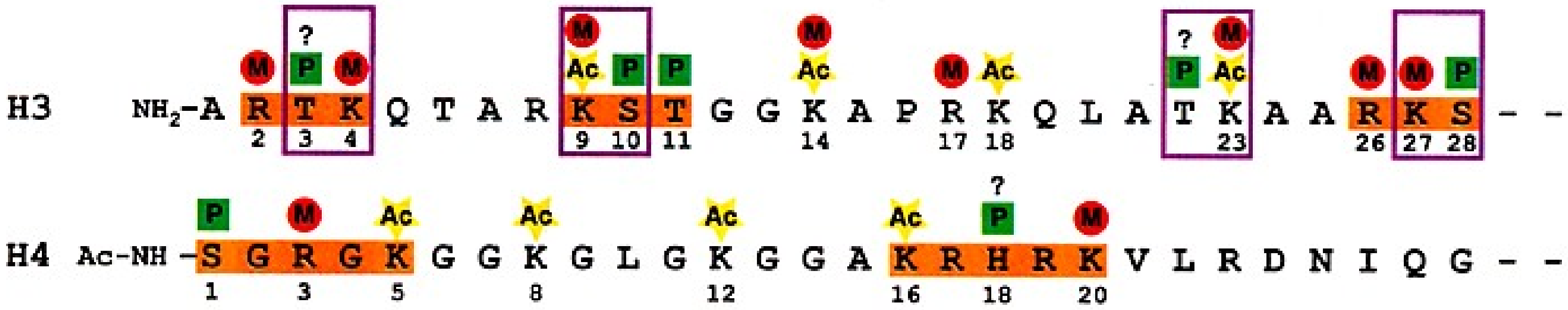
Alberts et al, Molecular Biology of the Cell (2002) Garland

Local binary switch: methyl/phos



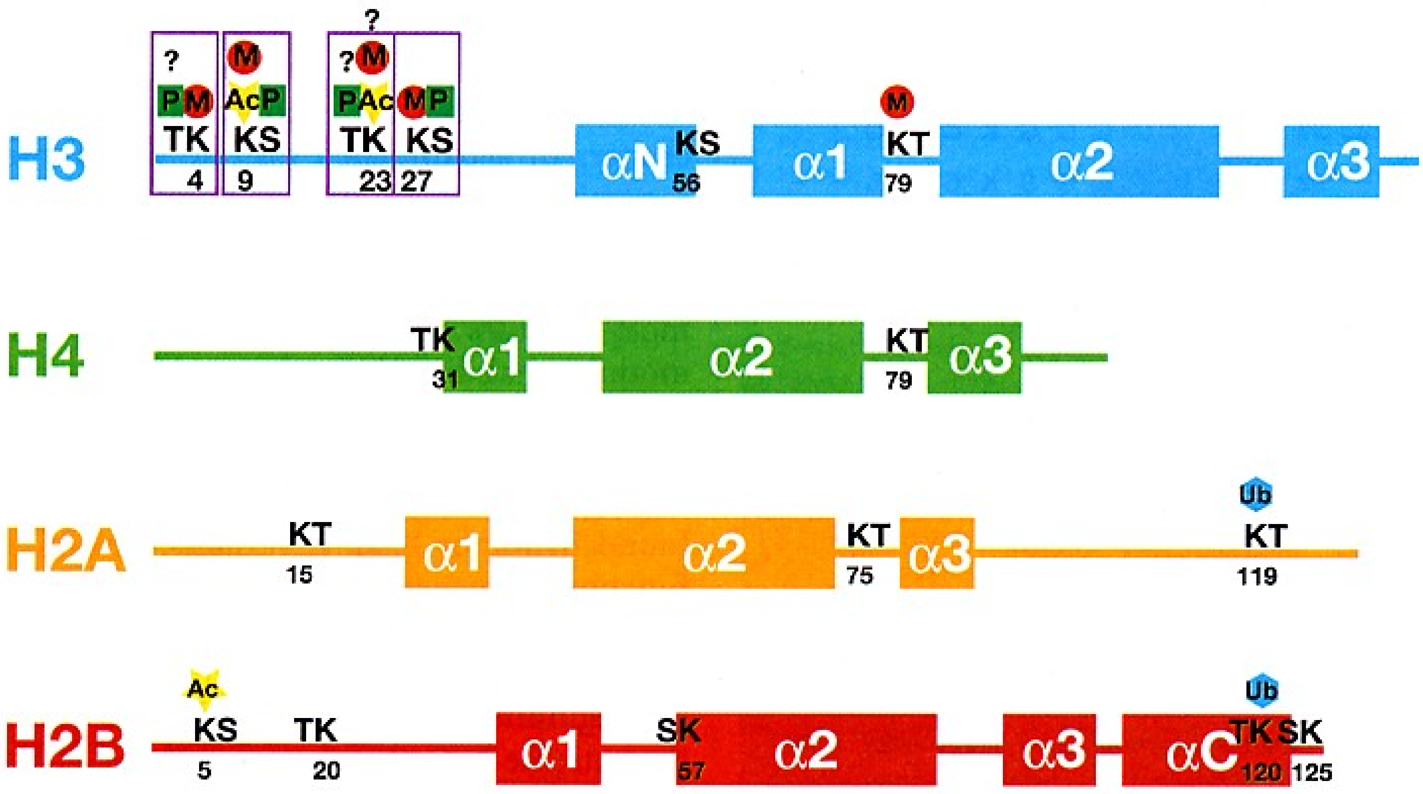
Fischle, Wang & Allis, Nature (2003) 425, 475

Clusters of histone marks


H3
NH2-A R T K Q T A R K S T G G K A P R K Q L A T K A A R K S - -
2 3 4 9 10 11 14 17 18 23 26 27 28
M P M ? M Ac P P M Ac M Ac ? P M Ac M M P
H4
Ac-NH -S G R G K G G K G L G K G G A K R H R K V L R D N I Q G - -
1 3 5 8 12 16 18 20
P M Ac Ac Ac Ac ? P M

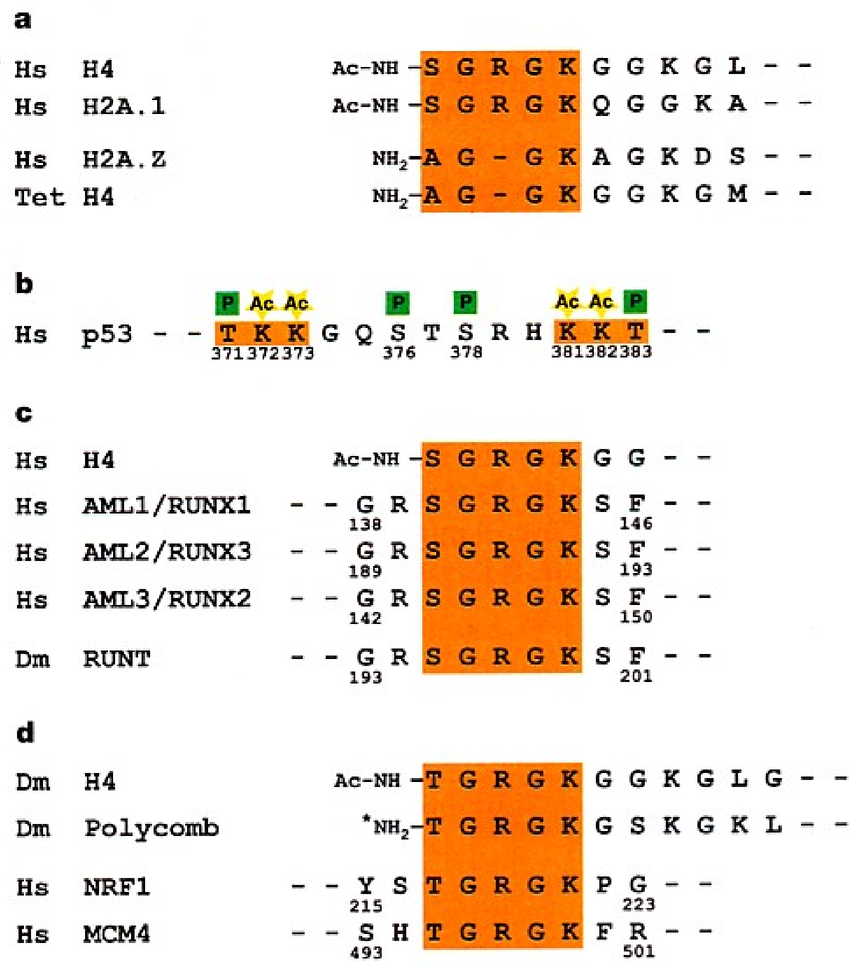
Fischle, Wang & Allis, Nature (2003) 425, 475

Putative methyl/phos switches



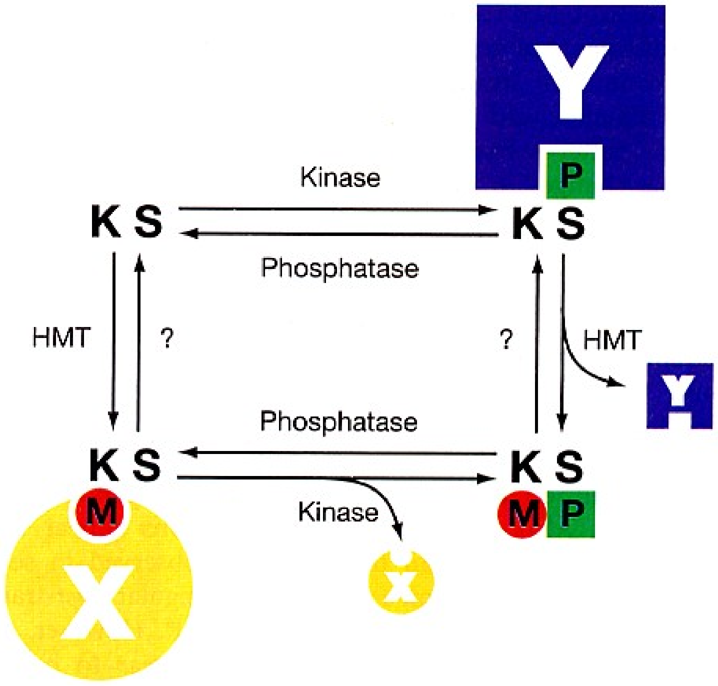
Fischle, Wang & Allis, Nature (2003) 425, 475

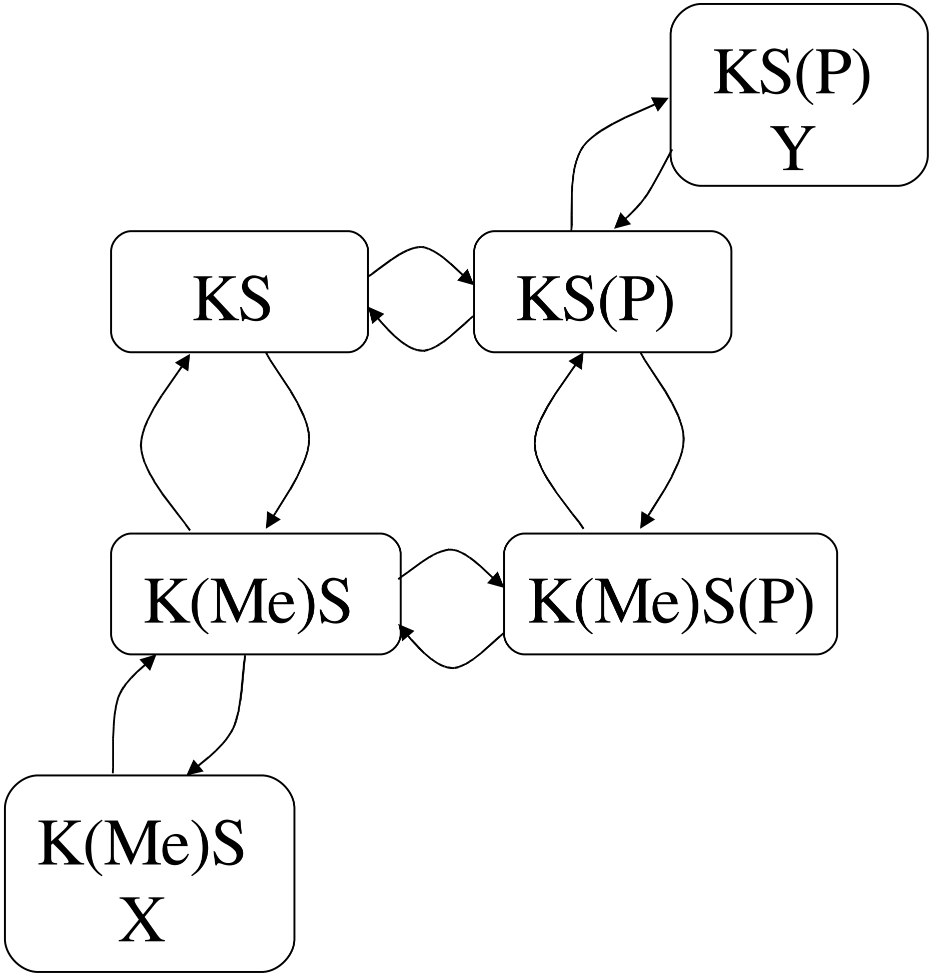
Putative switches in other proteins


a
Hs H4 Ac-NH -S G R G K G G K G L - -
Hs H2A.1 Ac-NH -S G R G K Q G G K A - -
Hs H2A.Z NH_2-A G - G K A G K D S - -
Tet H4 NH_2-A G - G K G G K G M - -
b
P Ac Ac P P Ac Ac P
Hs p53 - - T K K G Q S T S R H K K T - -
371 372 373 376 378 381 382 383
c
Hs H4 Ac-NH -S G R G K G G - -
Hs AML1/RUNX1 - - G R S G R G K S F - -
138 146
Hs AML2/RUNX3 - - G R S G R G K S F - -
189 193
Hs AML3/RUNX2 - - G R S G R G K S F - -
142 150
Dm RUNT - - G R S G R G K S F - -
193 201
d
Dm H4 Ac-NH -T G R G K G G K G L G - -
Dm Polycomb *NH_2-T G R G K G S K G K L - -
Hs NRF1 - - Y S T G R G K P G - -
215 223
Hs MCM4 - - S H T G R G K F R - -
493 501

Fischle, Wang & Allis, Nature (2003) 425, 475

Local binary switch: states




KS(P)
Y

KS

KS(P)

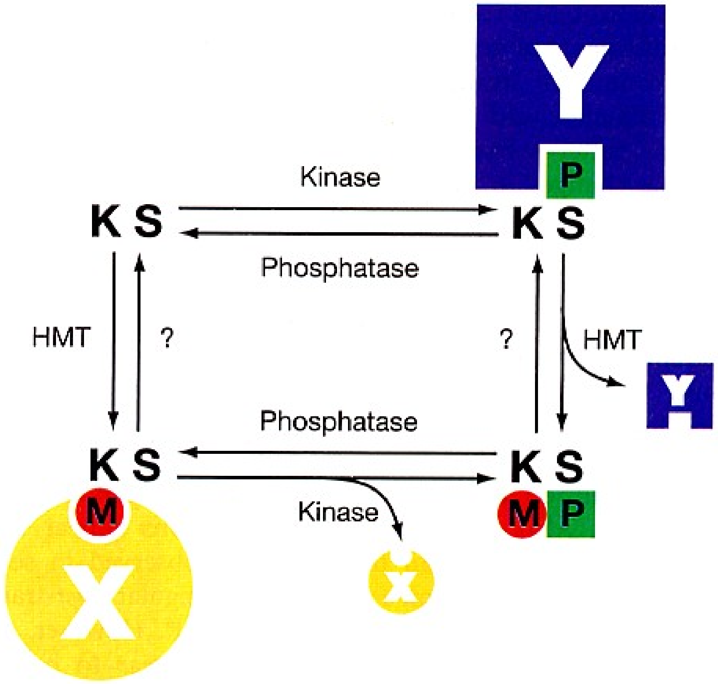
K(Me)S

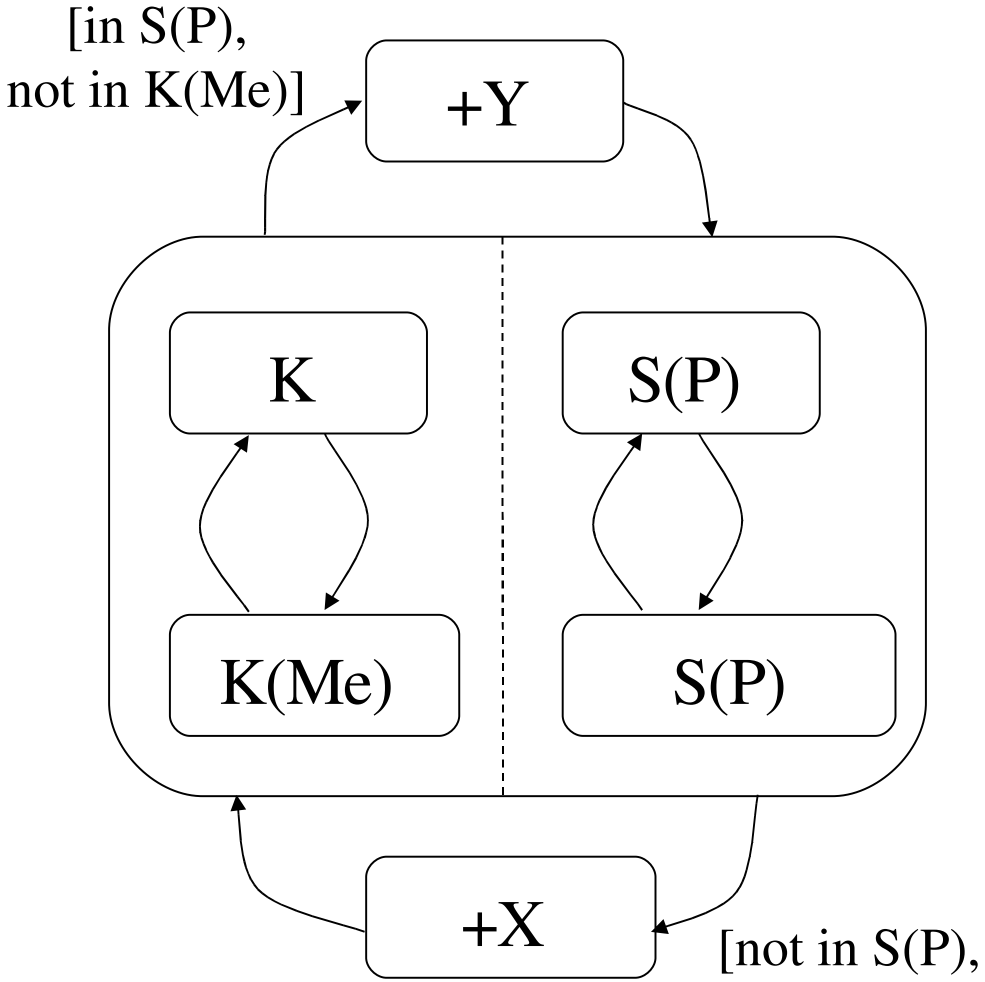
K(Me)S(P)

K(Me)S
X

Fischle, Wang & Allis,
Nature (2003) 425, 475

Local binary switch: orthogonality?





[in S(P),
not in K(Me)]

+Y

K

S(P)

K(Me)

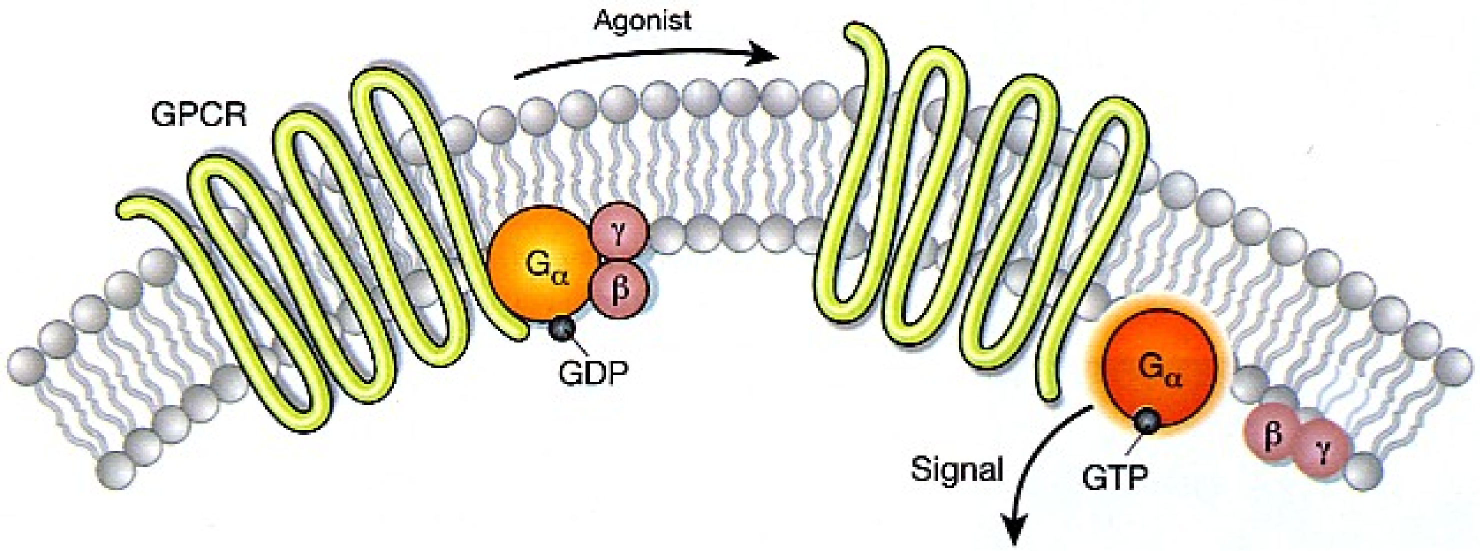
S(P)

+X

[not in S(P),
in K(Me)]

Fischle, Wang & Allis,
Nature (2003) 425, 475

Typical state change(s) in signaling



Li et al, Nature (2002) 420, 716

GeneCV issues: terms and concepts must be adapted

- Creation and destruction of proteins, complexes:
  - Central to biology
  - Not a primitive in Statecharts
- Hardwire some states?
  - Location
- Events (broadcast) irrelevant?

GeneCV issues: phenomenon vs mechanism

- States describing phenomena:
  - Active vs inactive
  - Binding vs non-binding
- States describing mechanism:
  - Phosphorylated or not
  - Folded vs unfolded
- How relate?
  - Combine to one state?
  - Relate two separate states?

GeneCV issues: object classes and inheritance

- Objects should belong to classes
- Should all objects be classes?
- Inherit properties from parent classes:
  - JNK1 is a tyrosine kinase
  - A tyrosine kinase is an ATP-dependent kinase
  - A kinase is an enzyme
  - An enzyme is a protein
- Classes for states and transitions?

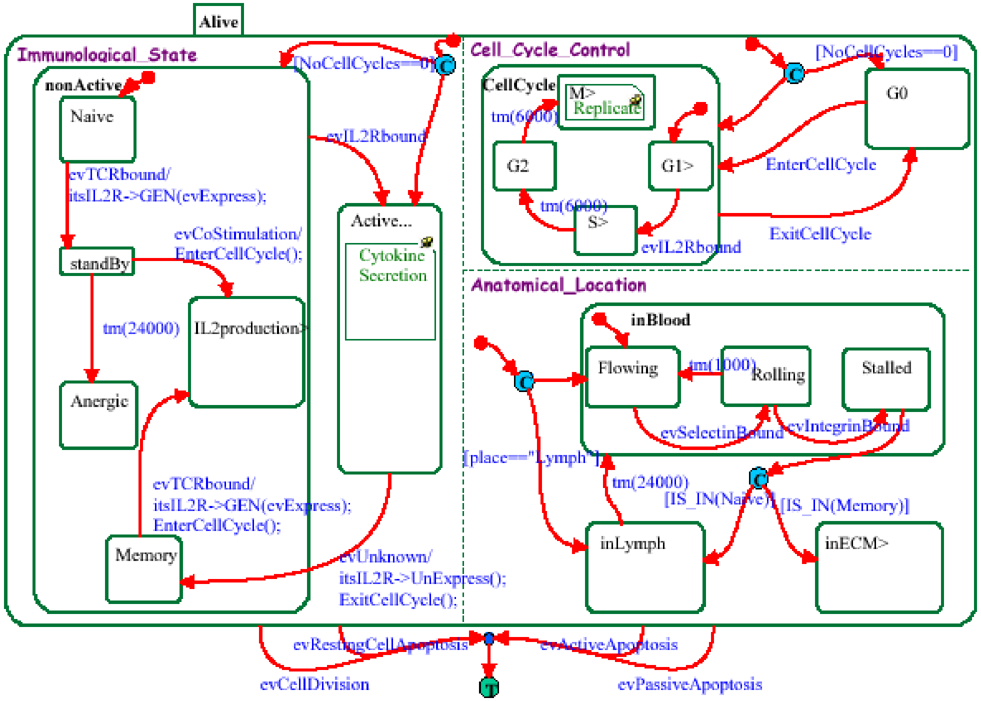
GeneCV issues:
state classes and inheritance?

- Classes for states (and transitions?)
- More powerful than generic states
- Ramifications?

GeneCV issues: implementation

- Graphics: same level as objects?
- Update and modification policies?
- Consistency checks
- Computational tools

GeneCV issues: extensions to higher levels



- How to extend to language of signaling: activation, inactivation…
- Biological life processes, scenarios

Kam, Cohen & Harel, VLFM’01 (2001)