The Abdus Salam
International Centre
for Theoretical Physics
50th Anniversary 1964–2014

2584-7

Spring College on the Physics of Complex Systems

26 May – 20 June, 2014

Statistical modeling of biological sequences

Martin Weigt
Université Pierre et Marie Curie
Paris

Strada Costiera, 11 - 34151 - Trieste - Italy • Tel. +39 0402240111 • Fax. +39 040224163 • sci_info@ictp.it • www.ictp.it
ICTP is governed by UNESCO, IAEA, and Italy, and it is a UNESCO Category 1 Institute

Statistical modeling of biological sequences

Martin Weigt

Computational and Quantitative Biology Department

Université Pierre et Marie Curie

Trieste

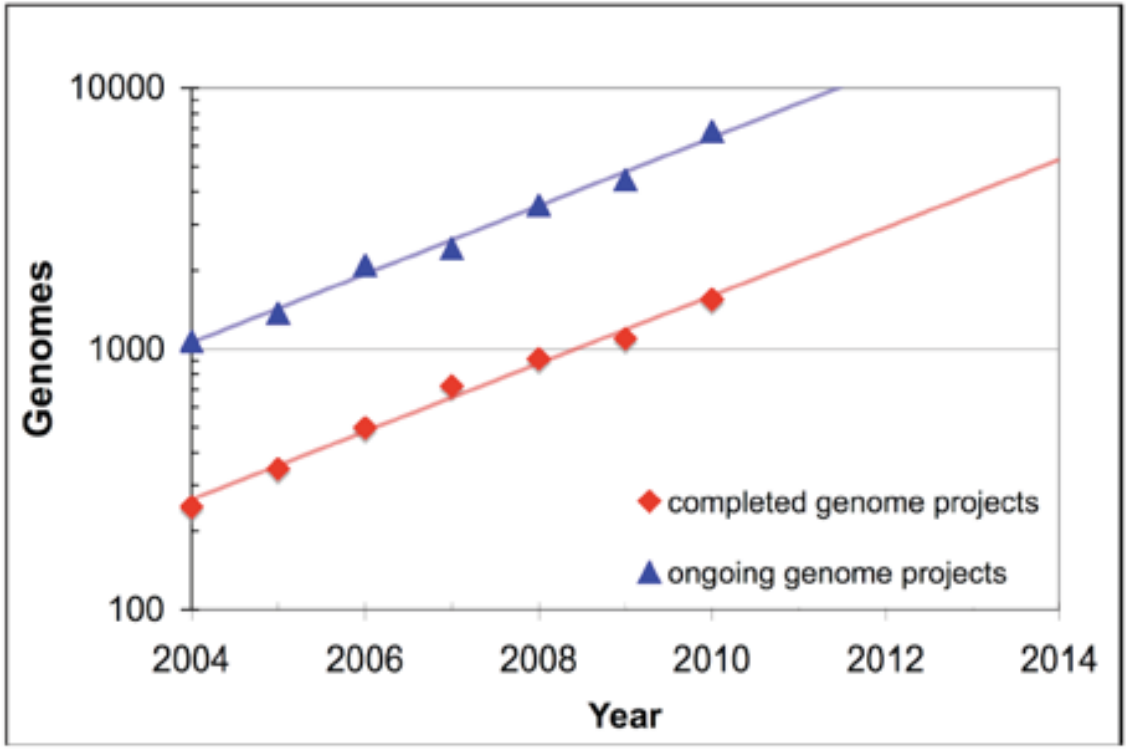
May 2014

Is there information in

?

There are many data...

- >19,500 completed genome sequencing projects
- >23,800 ongoing genome sequencing projects

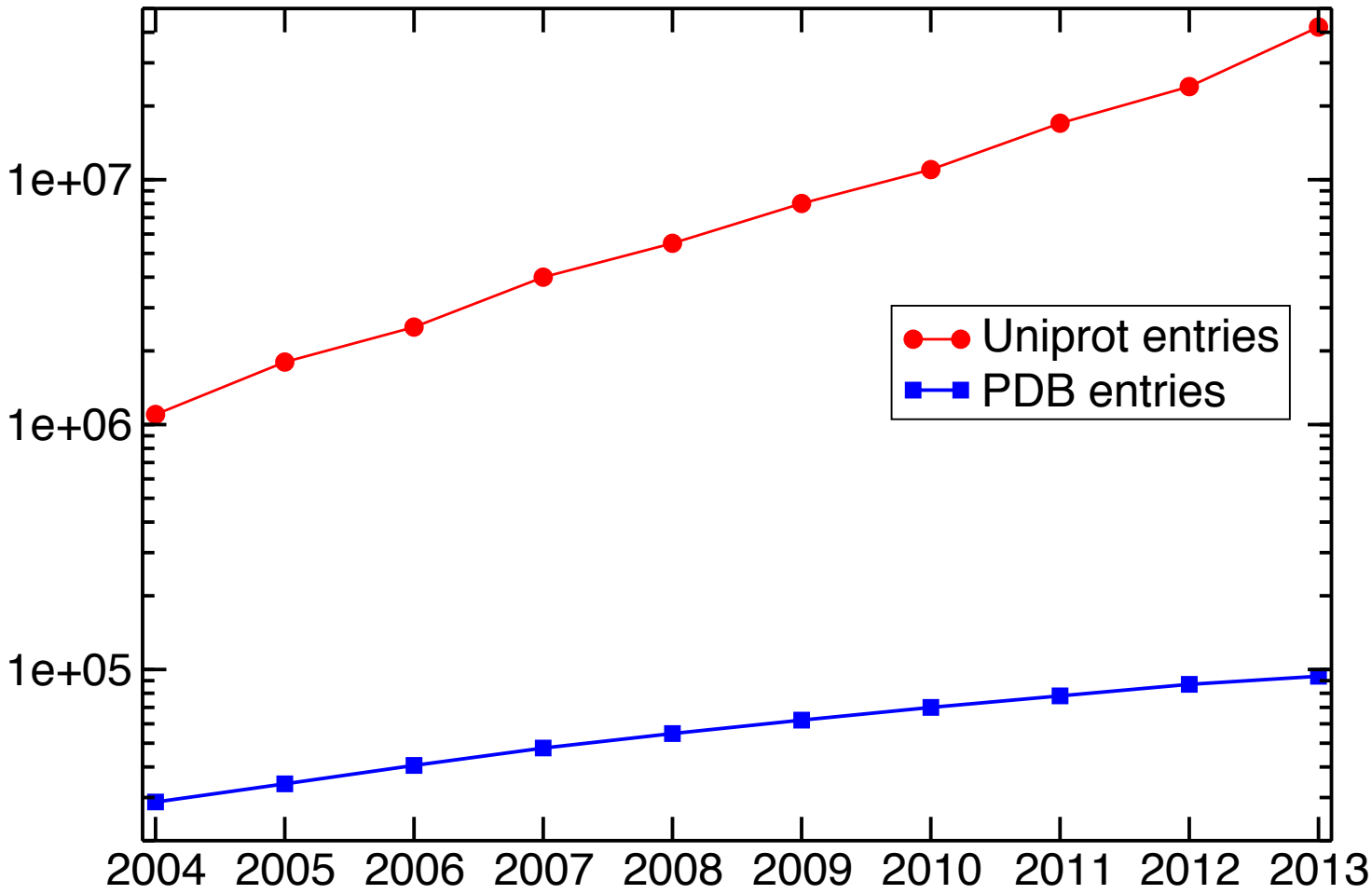


GOLD data base

- but genomes are long sequences of letters

➡ need computational approaches to extract information from raw data

Sequence vs. structure

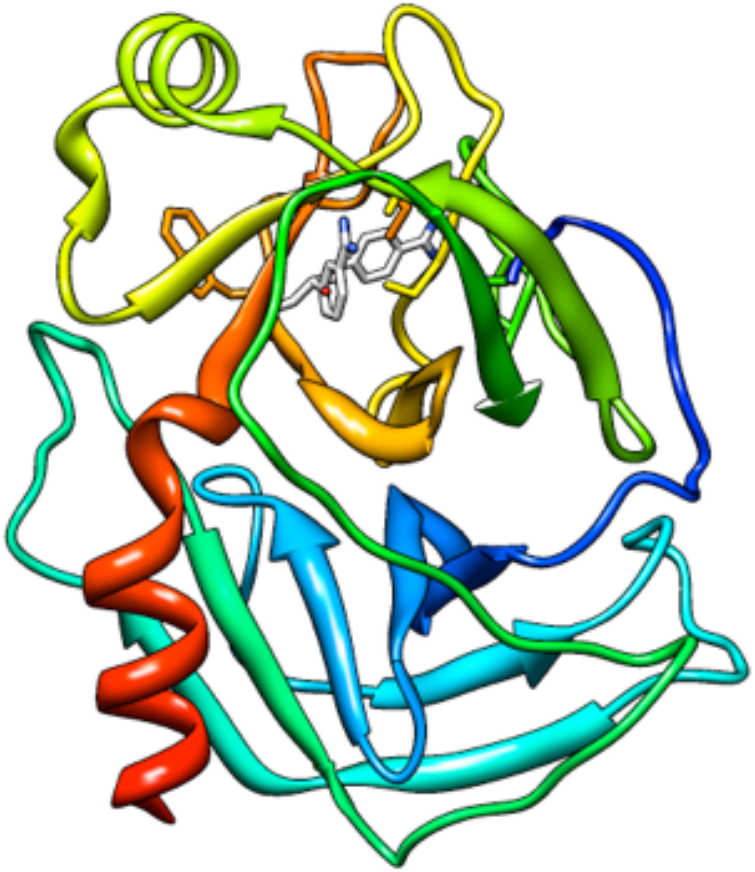


…but function relies on structure!

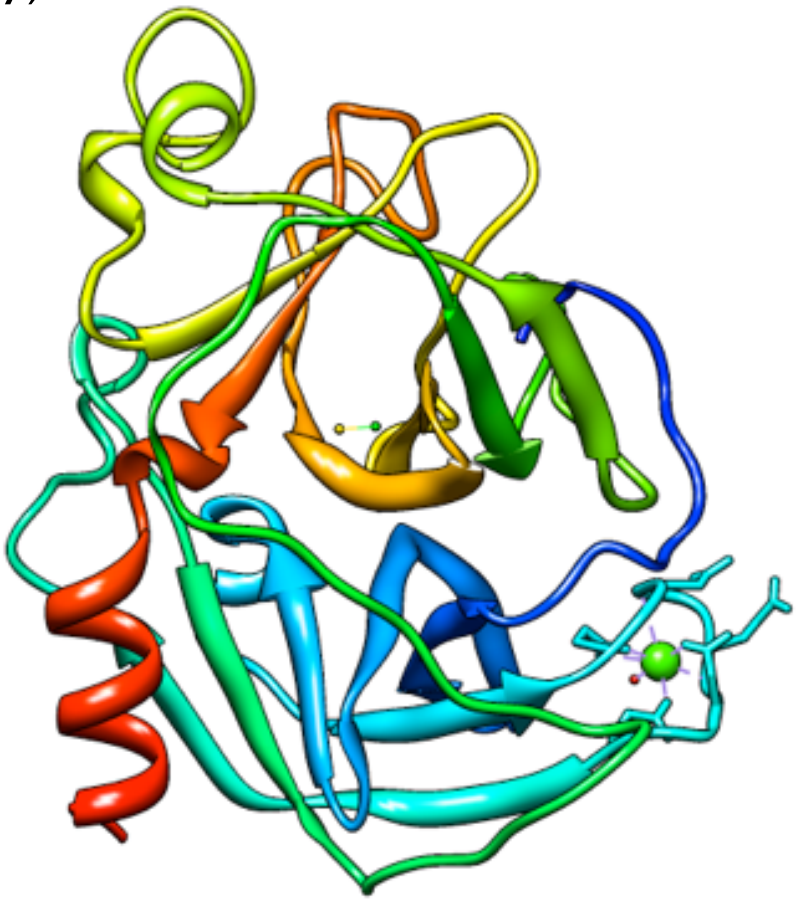
…and structure is more conserved in evolution than sequence

Structural conservation of protein domains

The structure domain folds in families frequently more conserved than amino-acid sequence (~25% sequence identity)



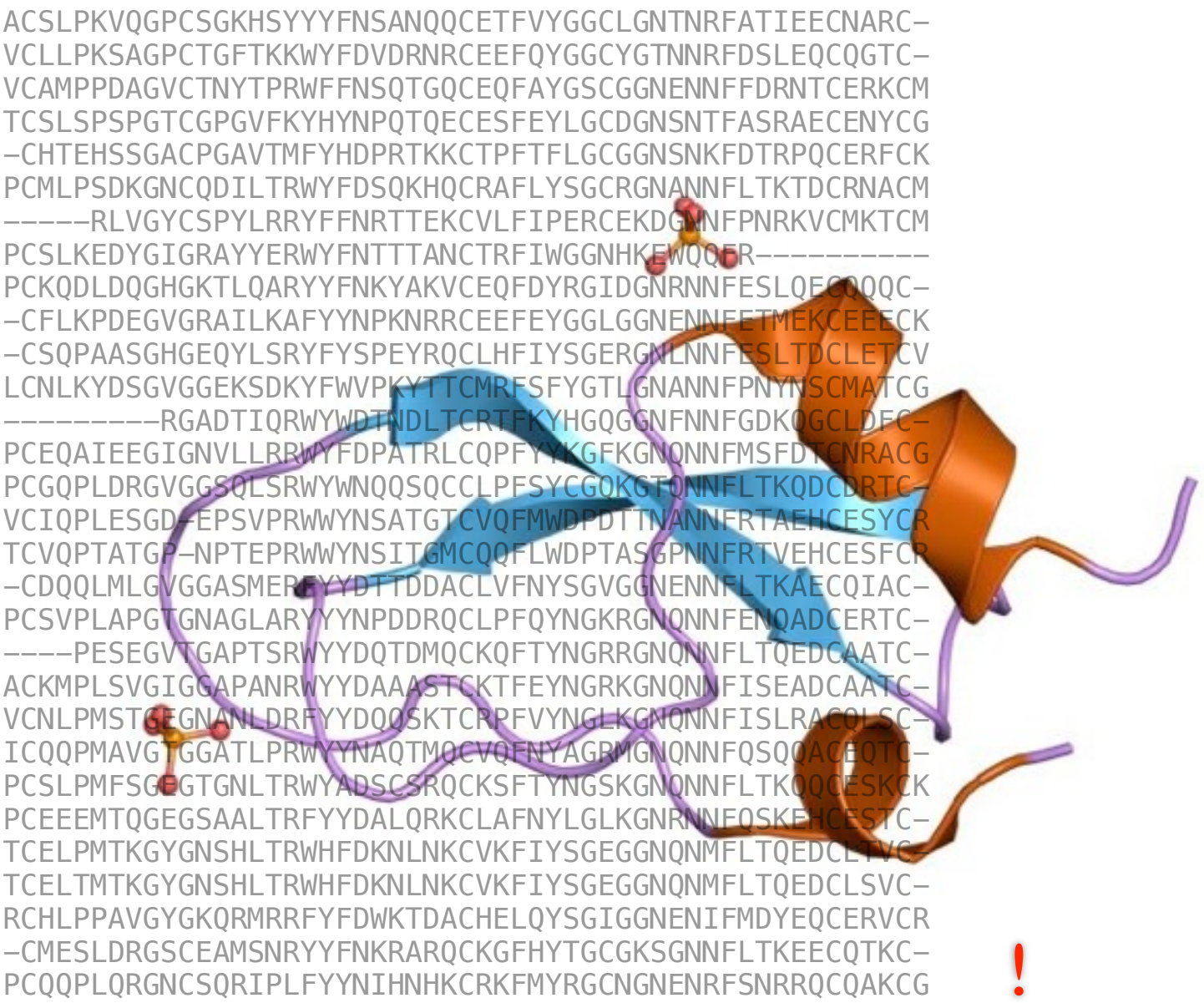
Trypsin (bovine)



Elastase (pig)

Sequence variability expected to carry information about structure

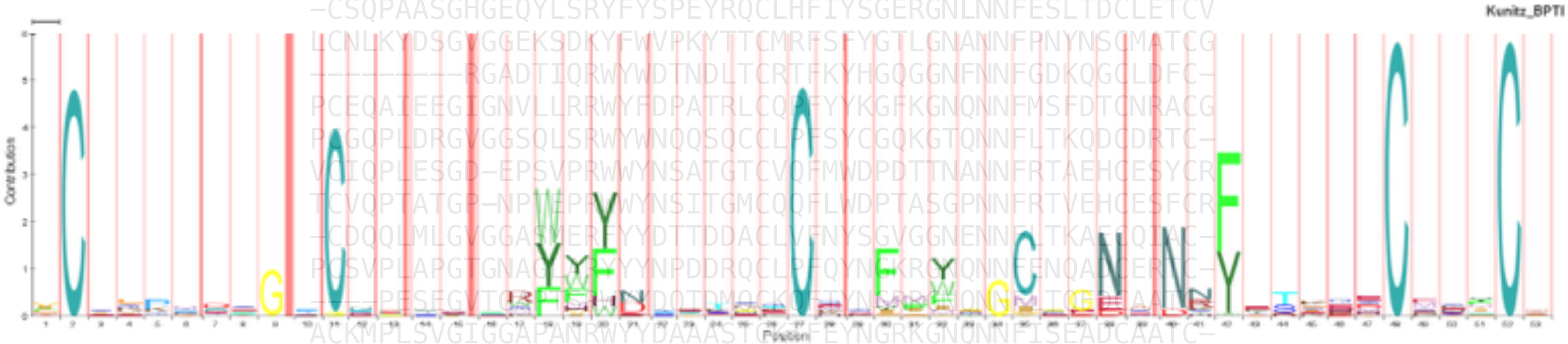
There is information in



!

There is information in

ACSLPKVQGPCSGKHSYYYFNSANQQCETFVYGGCLGNTNRFATIEECNARC-
VCLLPKSAGPCTGFTKKWYFDVDRNRCEEFQYGGCYGTNNRFDSLEQCQGTC-
VCAMPPDAGVCTNYTPRWFFNSQTGQCEQFAYGSCGGNENNFFDRNTCERKCM
TCSLSPSPGTCGPGVFKYHYNPQTQECESFEYLGCDGNSNTFASRAECENYCG
-CHTEHSSGACPGAVTMFYHDPRTKKCTPFTFLGCGGNSNKFDTRPQCERFCK
PCMLPSDKGNCQDILTRWYFDSQKHQCRAFLYSGCRGNANNFLTKTDCRNACM
-----RLVGYCSPYLRRYFFNRTTEKCVLFIPERCEKDGNNFPNRKVCMKTCM
PCSLKEDYGIGRAYYERWYFNTTTANCTRFIWGGNHKEWQQFR----------
PCKQDLDQGHGKTLQARYYFNKYAKVCEQFDYRGIDGNRNNFESLQECQQQC-
-CFLKPDEGVGRAILKAFYYNPKNRRCEEFEYGGLGGNENNFETMEKCEEECK
-CSQPAASGHGEQYLSRYFYSPEYRQCLHFIYSGERGNLNNFESLTDCLETCV
LCNLKYDSGVGGEKSDKYFWVPKYTTCMRFSFYGTLGNANNFPNYNSCMATCG
---------RGADTIQRWYWDTNDLTCRTFKYHGQGGNFNNFGDKQGCLDFC-
PCEQAIEEGIGNVLLRRWYFDPATRLCQPFYYKGFKGNQNNFMSFDTCNRACG
PCGQPLDRGVGGSQLSRWYWNQQSQCCLPFSYCGQKGTQNNFLTKQDCDRTC-
VCIQPLESGD-EPSVPRWWYNSATGTCVQFMWDPDTTNANNFRTAEHCESYCR
TCVQPTATGP-NPWEPRWWYNSITGMCQQFLWDPTASGPNNFRTVEHCESFCR
-CDQQLMLGVGGASMERFYYDTTDDACLVFNYSGVGGNENNFLTKAECQIAC-
PCSVPLAPGTGNAGLARYYYNPDDRQCLPFQYNGKRGNQNNFENQADCERTC-
ACKMPLSVGIGGAPANRWYYDAAASTCKTFEYNGRKGNQNNFISEADCAATC-
VCNLPMSTGEGNANLDRFYYDQQSKTCRPFVYNGLKGNQNNFISLRACQLSC-
ICQQPMAVGTGGATLPRWYYNAQTMQCVQFNYAGRMGNQNNFQSQQACEQTC-
PCSLPMFSGEGTGNLTRWYADSCSRQCKSFTYNGSKGNQNNFLTKQQCESKCK
PCEEEMTQGEGSAALTRFYYDALQRKCLAFNYLGLKGNRNNFQSKEHCESTC-
TCELPMTKGYGNSHLTRWHFDKNLNKCVKFIYSGEGGNQNMFLTQEDCLTVC-
TCELTMTKGYGNSHLTRWHFDKNLNKCVKFIYSGEGGNQNMFLTQEDCLSVC-
RCHLPPAVGYGKQRMRRFYFDWKTDACHELQYSGIGGNENIFMDYEQCERVCR
-CMESLDRGSCEAMSNRYYFNKRARQCKGFHYTGCGKSGNNFLTKEECQTKC-
PCQQPLQRGNCSQRIPLFYYNIHNHKCRKFMYRGCNGNENRFSNRRQCQAKCG



!

Plan of the lectures

1. DNA sequence motifs, transcription-factor binding sites and position-specific weight matrices

   [van Nimwegen, BMC Bioinformatics (2007)]

2. Aligning biological sequences and detecting sequence similarity

   [Durbin, Eddy, Krogh, Mitchison, Biological Sequence Analysis, Cambridge 1998]

3. RNA secondary structure prediction

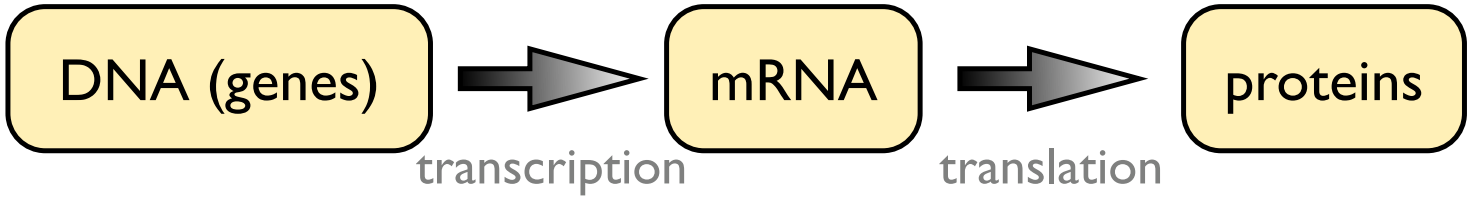
   [Durbin, Eddy, Krogh, Mitchison, Biological Sequence Analysis, Cambridge 1998]

4. Direct-coupling analysis: From sequence coevolution to protein structure

   [Morcos et al., PNAS (2011); Juan, Pazos, Valencia, Nature Rev Gen (2013)]

Gene regulation

Central dogma of molecular biology: directed information flow

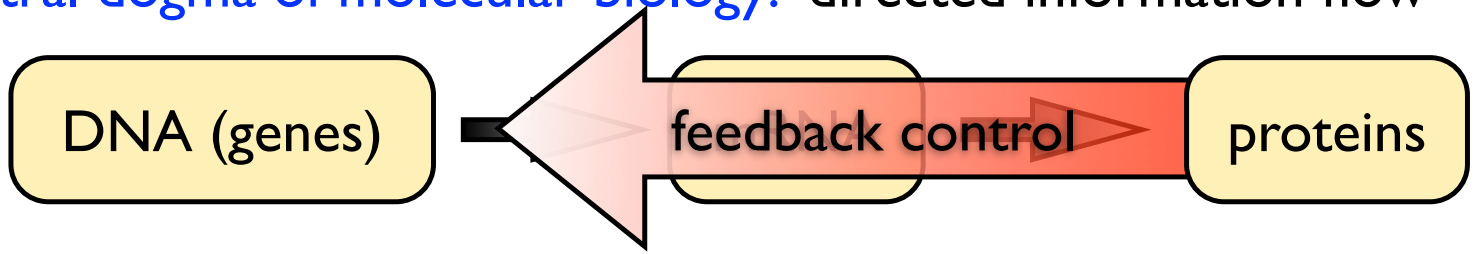


BUT

- different cell types from same genome
  - differential gene expression
- precise timing of gene expression during cell cycle
- response to external signals, nutrient availability etc.

Gene regulation

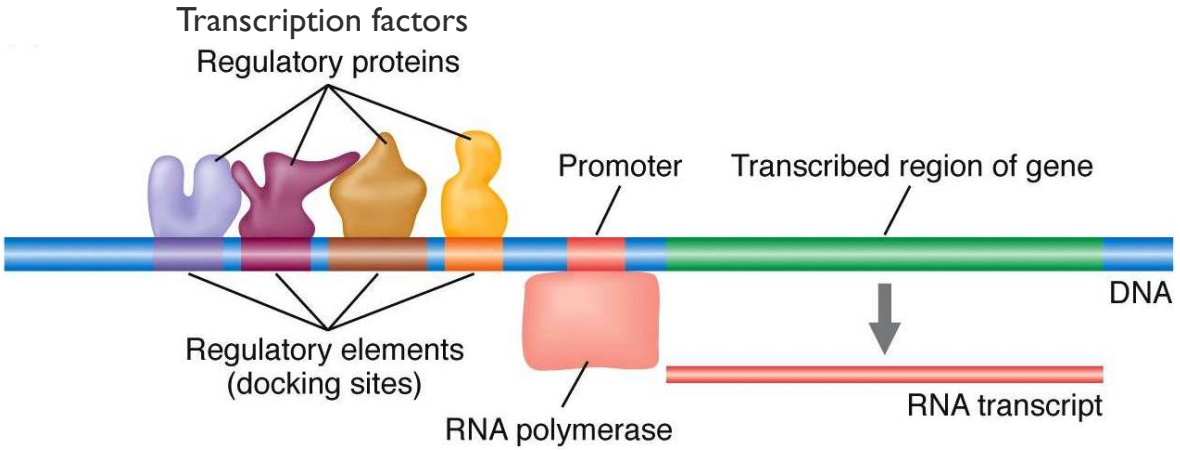
Central dogma of molecular biology: directed information flow



BUT

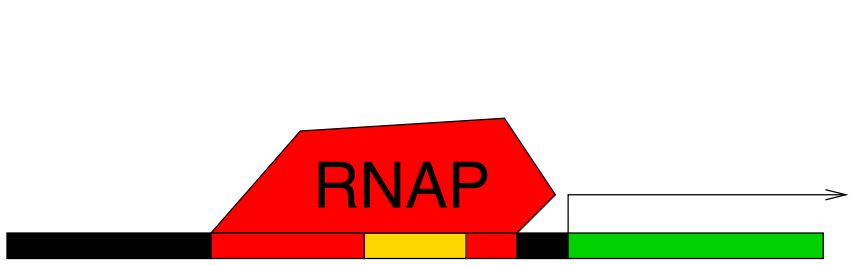
- different cell types from same genome
  - differential gene expression
- precise timing of gene expression during cell cycle
- response to external signals, nutrient availability etc.

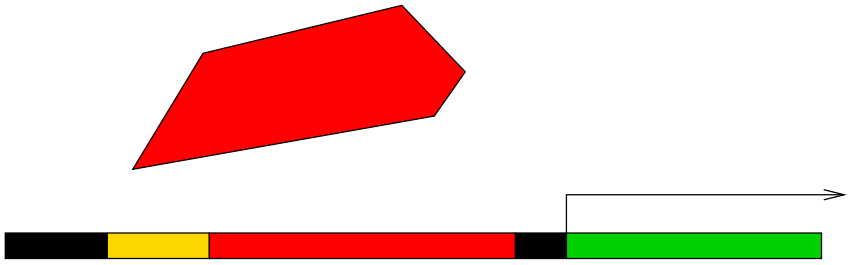
Gene regulation = fundamental process for differential gene expression

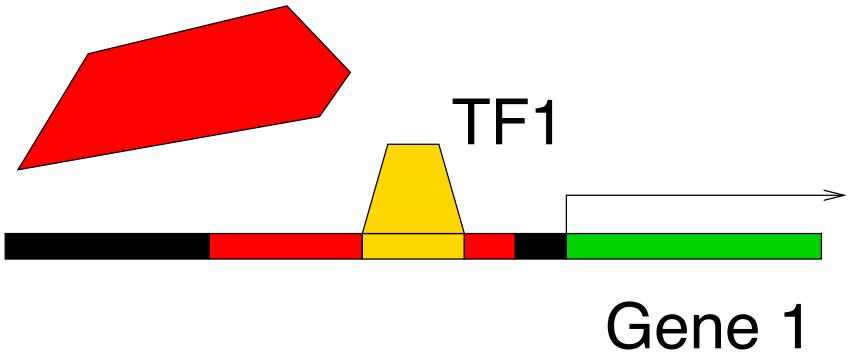


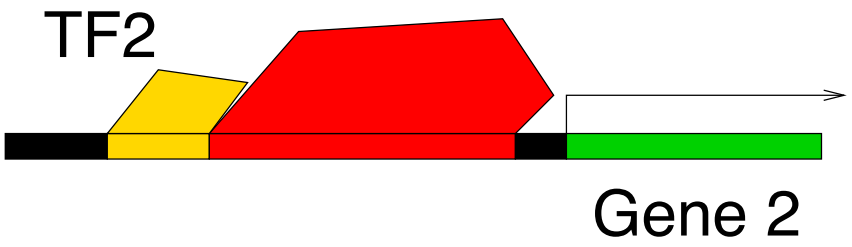
Transcriptional repression vs. activation

Simplest regulatory functions:







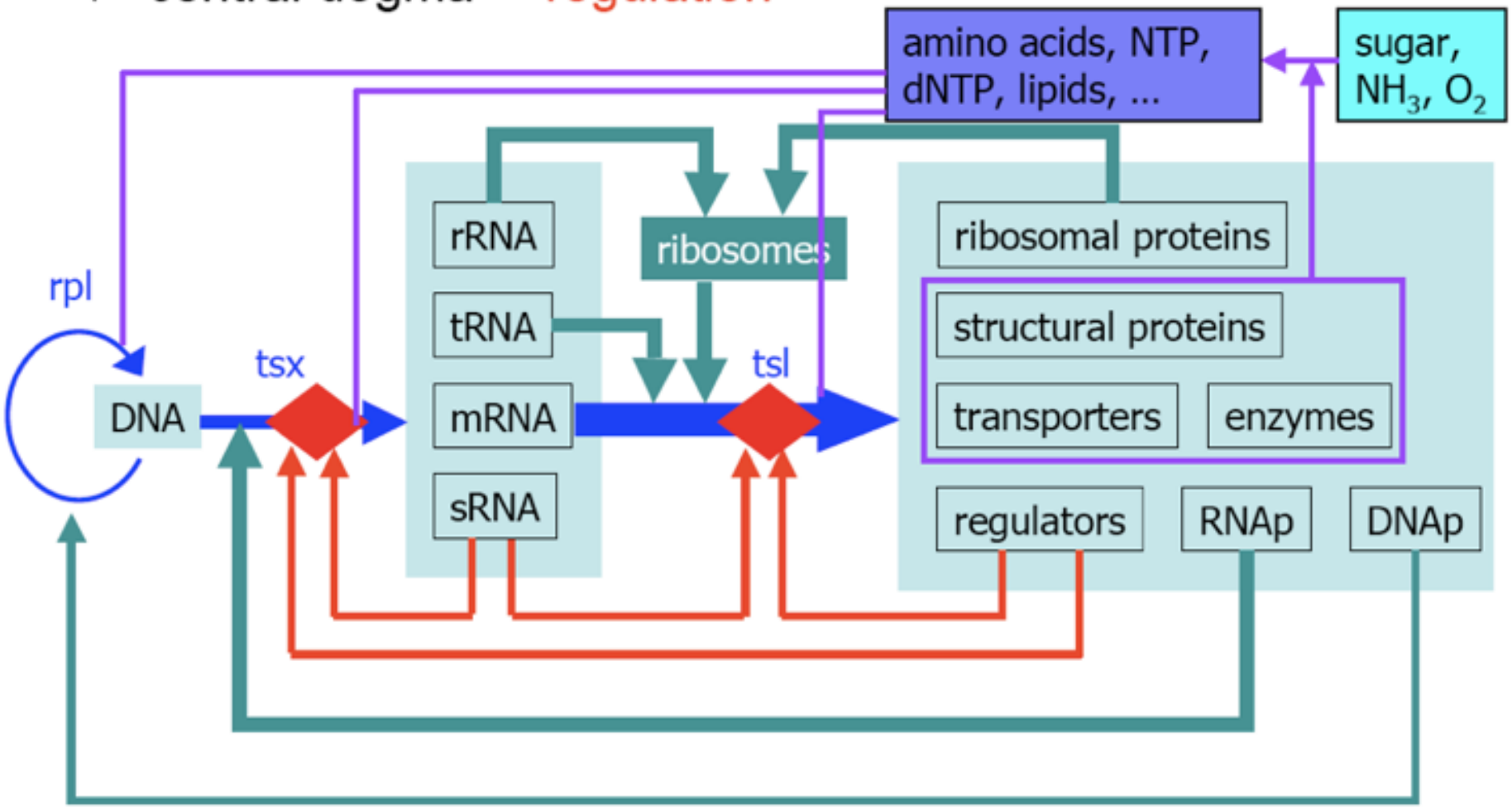


repression
by competitive binding

activation
by collective binding

Gene regulation

❖ central dogma + regulation

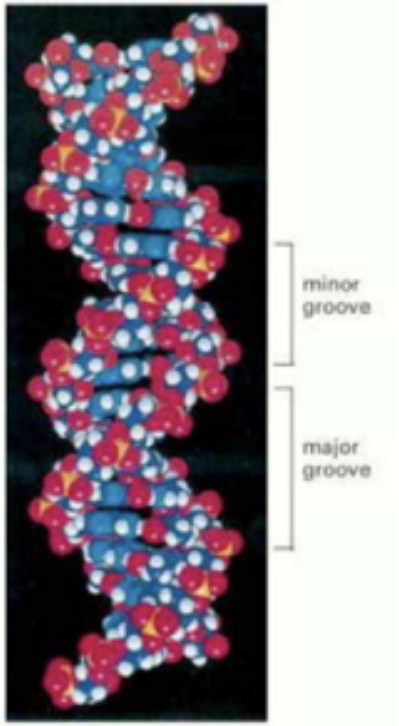


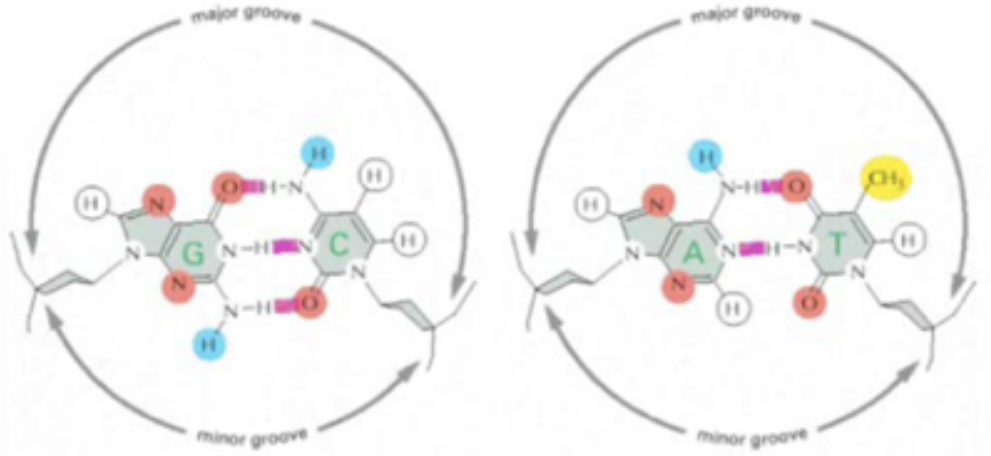
➡ concentrate on transcriptional regulation

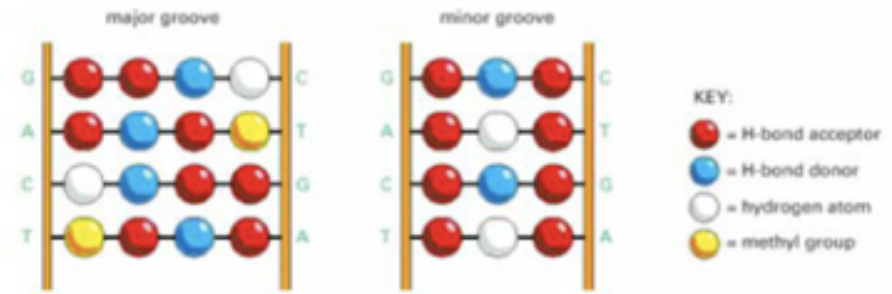
Protein-DNA interactions

A. Empirical facts

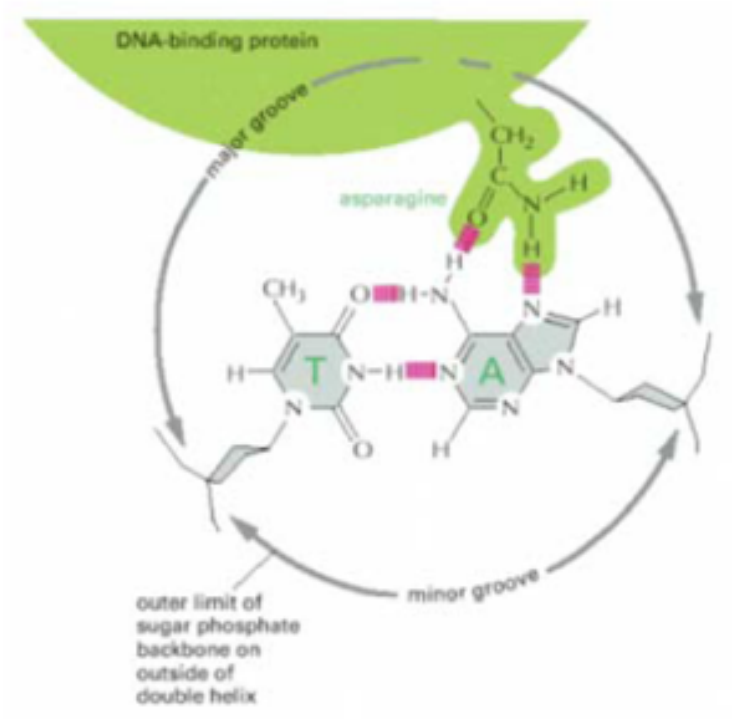
1. Transcription Factors
   - size: ~5nm (10-20 bp)
   - molecular basis of sequence recognition

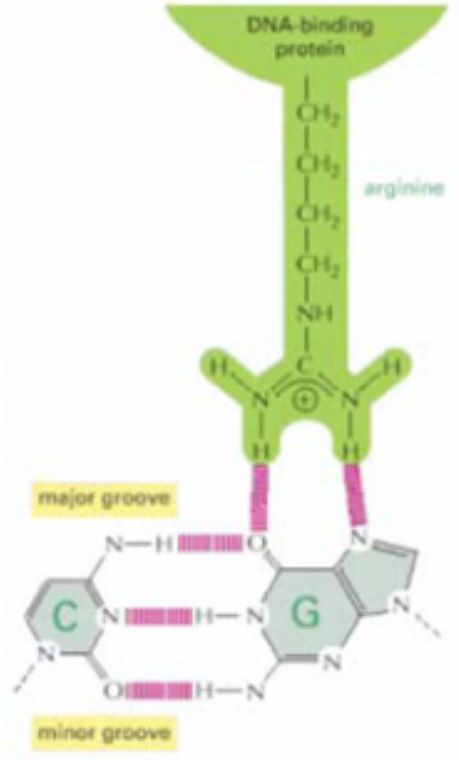






- contact between TF and DNA





➔ structure of a TF must place the appropriate amino acids next to the base pairs they contact

2. DNA binding sequences

- typically 10-20 bp in bacteria

protein	target sequence
lac repressor	5' AATTGTGAGCGGATAACAATT 3' TTAACACTCGCCTATTGTTAA
CRP	TGTGAGTTAGCTCACT ACACTCAATCGAGTGA
λ repressor	TATCACCGCCAGAGGTA ATAGTGGCGGTCTCCAT

- lots of sequence variants
- consensus sequence often palindromic
- common to have 2~3 mismatches from the core consensus sequence -- "fuzzy" binding motif

3. TF-DNA interaction

- passive (no energy consumption)
- strong electrostatic attraction indept of binding seq

  e.g., $[TF-DNA] > 10 \times [TF]_{free}$ for LacI in 0.1M salt

  ➔ non-specific binding: $G_{ns} - G_{cyto} \simeq -15kT$

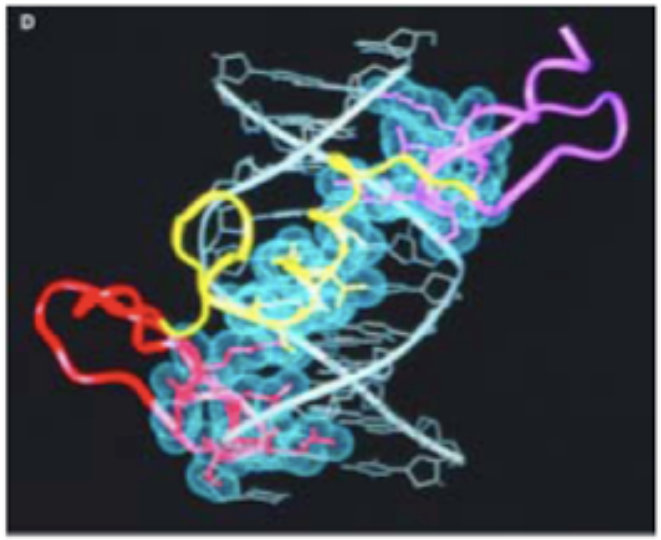
  ($kT \approx$ 0.62 kcal/mole at 37C)

- additional energy gained from hydrogen bonds to preferred sequences

  strongest binder: $G^* - G_{ns} \simeq -15kT$

- graded increase in binding energy for sequences with partial match to the preferred sequence

- relative binding affinity for Mnt



binding energy matrix

(in unit of kT ≈ 0.6 kcal/mole)

pos.	10	11	12	13	14	15	16	17
A	1.8	2.4	1.6	1.0	0	2.1	0.8	1.1
C	2.4	1.9	4.2	2.1	0.3	0	0	0
G	0	1.6	0	0	1.2	3.2	1.0	1.2
T	3.0	0	2.2	2.2	0.6	2.2	0.7	0.3

(D.S. Fields, Y. He, A. Al-Uzri & G. Stormo, 1997)

(from competitive binding expts)

➔ weak energetic preference -- weak specificity

➔ similar results for other TFs studied (e.g., LacI, λ-CI, λ-Cro)

- double mutation: binding energy approx additive

➔ Can we say something generic about the design of TF-DNA interaction from these facts/data?