

Comparative Genomics

of Regulatory Signals

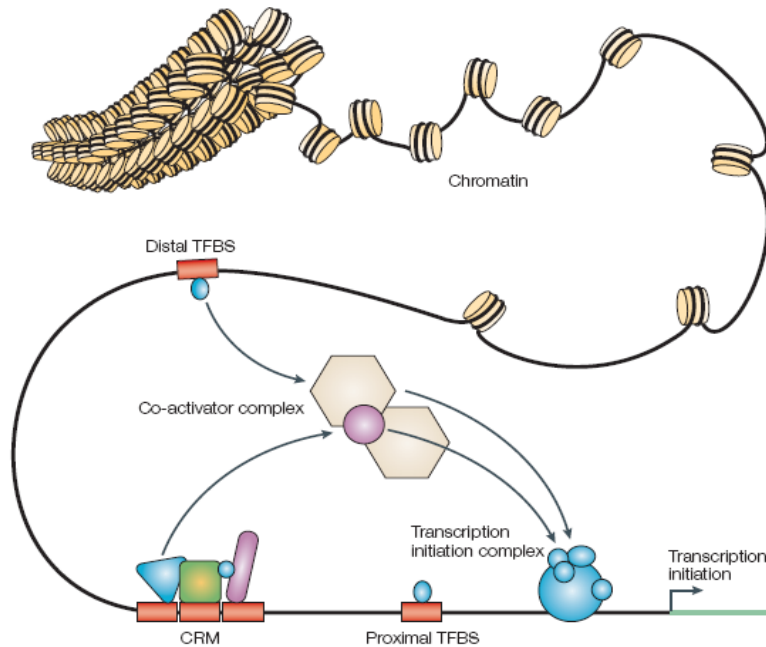
Outline

- I. Introduction
- II. Biophysics of regulation
- III. Finding regulatory elements
- IV. Annotation of signals
- V. Evolution of regulation

Introduction

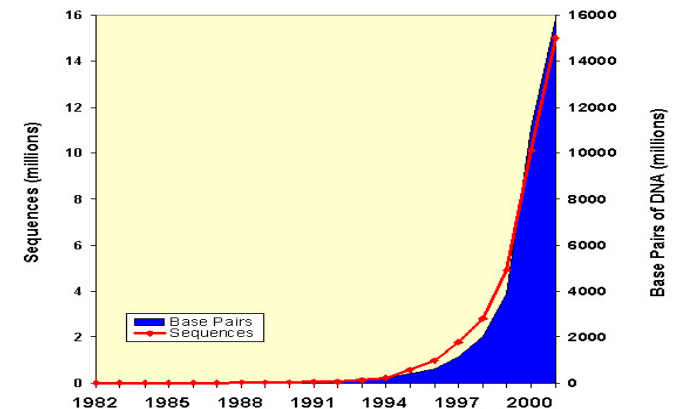
- Current genomics

Deciphering regulatory control mechanisms that govern gene expression



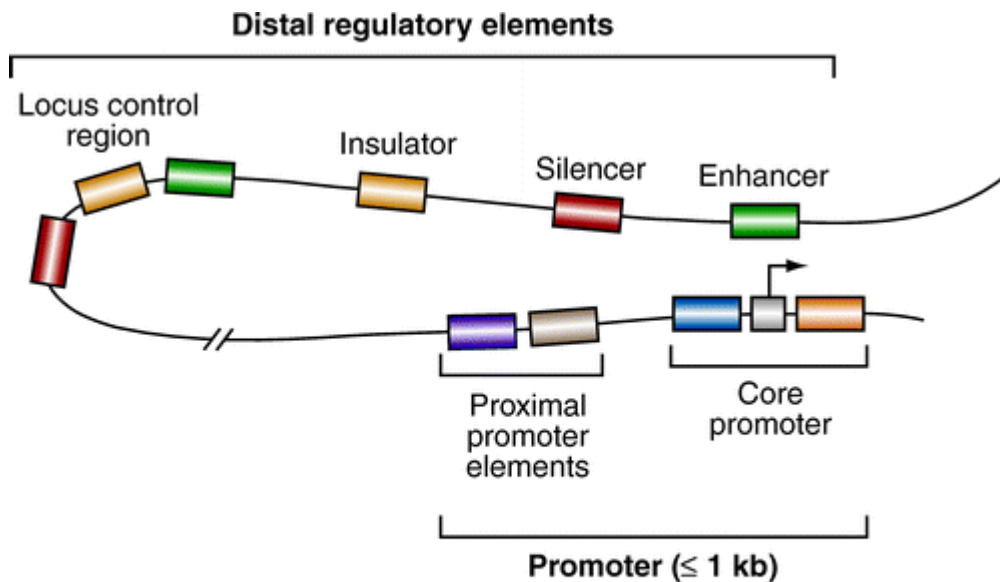
Components of transcriptional regulation


Growth of GenBank



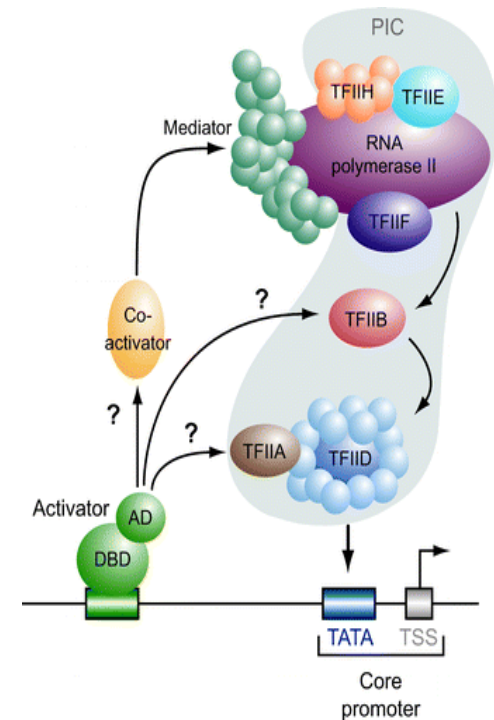
Regulatory apparatus


- Cis-elements –promoters, enhancers (TFBS)
- Trans-elements-transcription factors



 Maston GA, et al. 2006.
Annu. Rev. Genomics Hum. Genet. 7:29–59

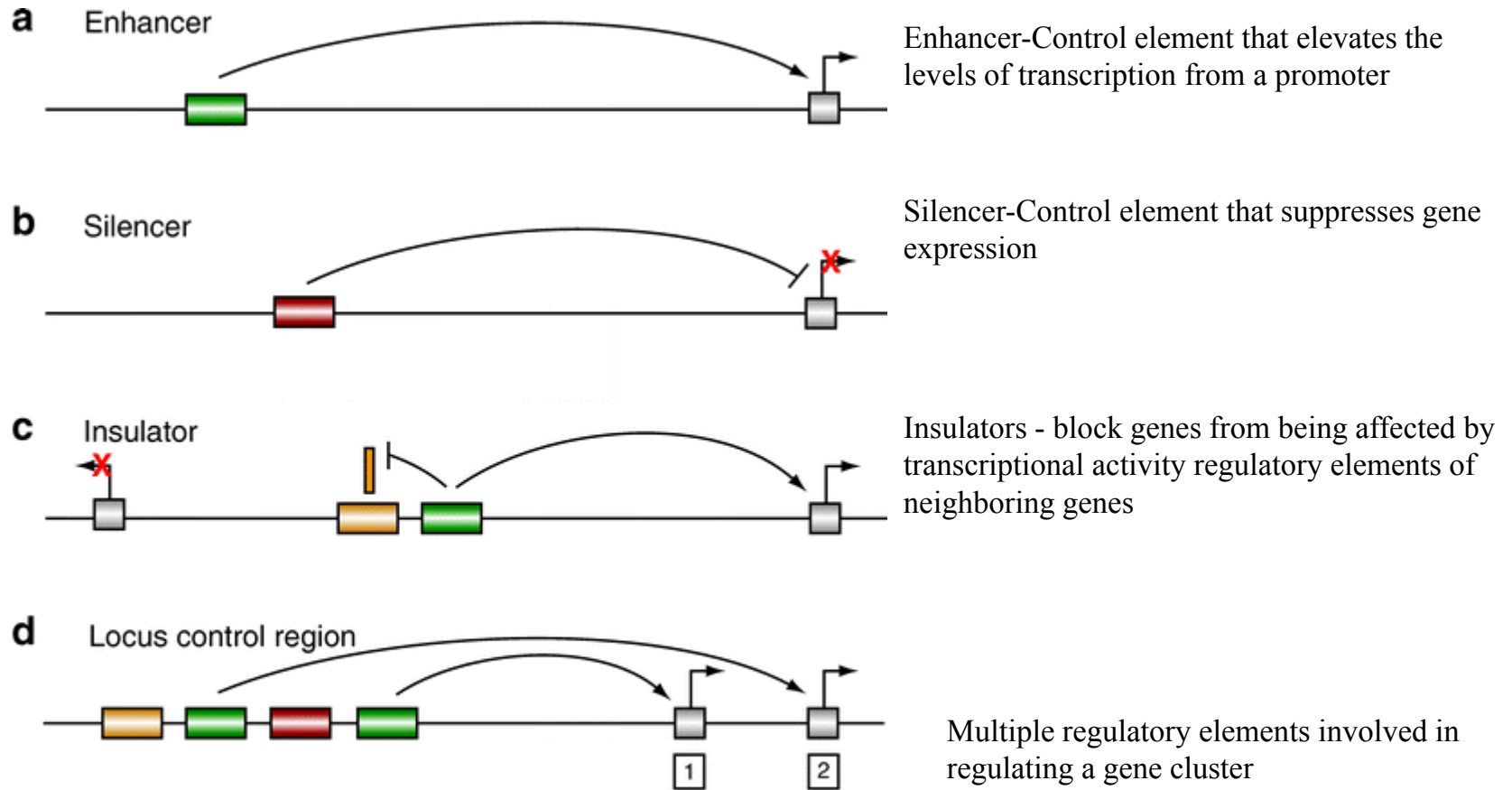
Schematic figure of a typical gene regulatory region.



 Maston GA, et al. 2006.
Annu. Rev. Genomics Hum. Genet. 7:29–59

The eukaryotic transcriptional machinery

Cis-regulatory elements



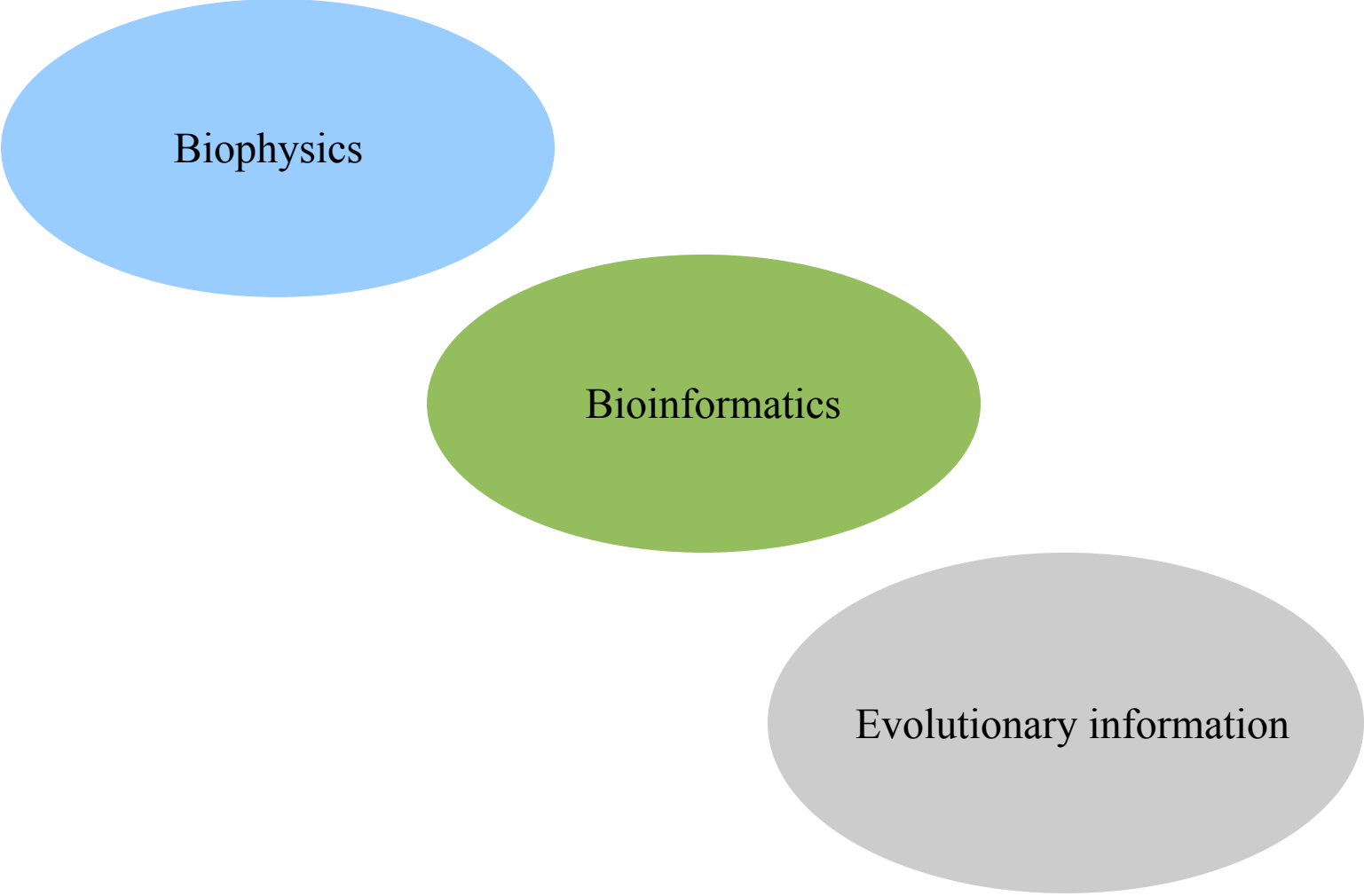
Identification of regulatory regions

- Identification of TATA-box sequences- ~30bp upstream transcription start site
- CpGs islands – methylation

Problems:

- Not all transcription-sites are proximal to CpG islands and the association between CpG and promoters is not present in all organisms

Making sense out of regulatory sequence data



Biophysics

Bioinformatics

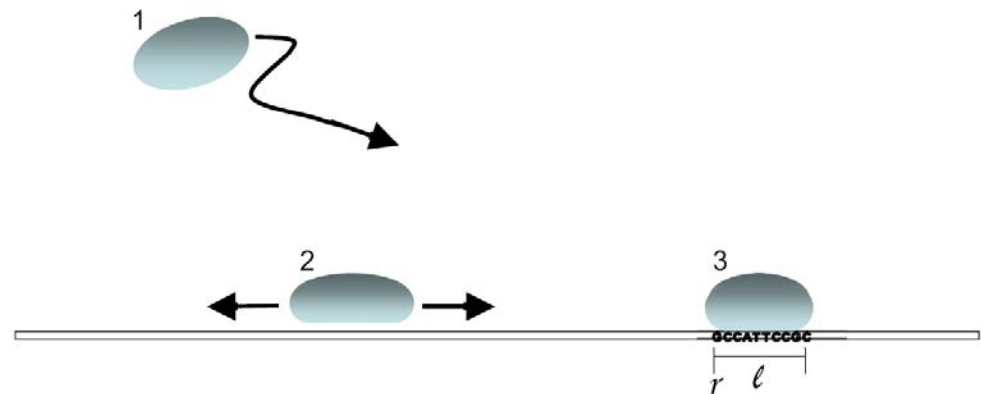
Evolutionary information

II - Biophysics of regulation

- Binding of a transcription factor
 - Binding energies
 - Example in *E. coli*
 - Search kinetics
- Thermodynamics of factor binding
 - Deriving probabilities
 - Bounds on genomic design of regulation
- Implications

Binding of a transcription factor

- 3 thermodynamic states
 1. Unbound
 2. Unspecific bound state (electrostatic interactions)
 3. Specific bound state (hydrogen bonds)



Binding of a transcription factor

- Binding energy
 - independent, additive contributions of single nucleotides in sequence $\mathbf{a} \equiv (a_1, \dots, a_l)$

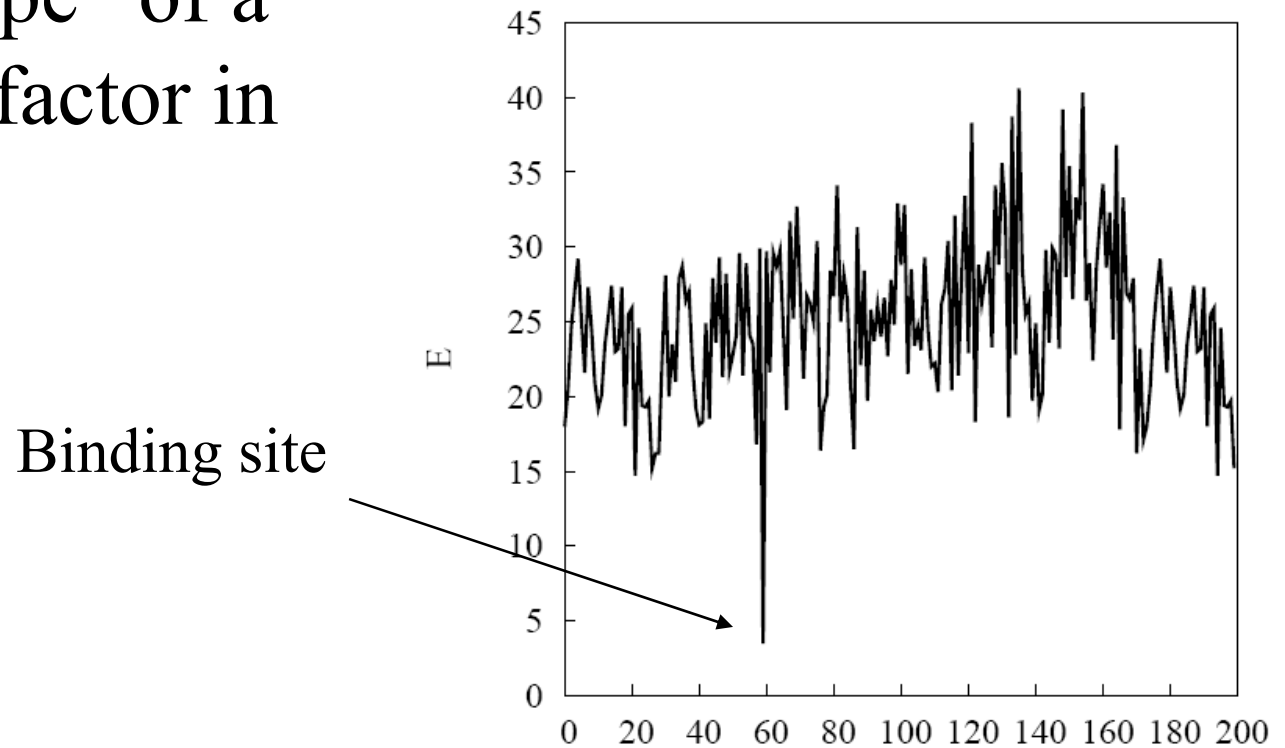
$$E(\mathbf{a}) = \sum_{i=1}^l \epsilon_i(a_i) \quad E^* \equiv E(\mathbf{a}^*)$$

- 2 state approximation: Binding energy simply related to Hamming distance and $\epsilon \approx 2k_B T$

$$E(\mathbf{a}) = E^* + \epsilon \cdot d_H(\mathbf{a}, \mathbf{a}^*)$$

Binding of a transcription factor

- Example for an energy „landscape“ of a specific factor in *E. coli*



Binding of a transcription factor

- Remarkably fast in the cell
- Search process modelled as a mixture between
 - 3D diffusion in medium („hopping“)
 - 1D diffusion along DNA backbone
- Kinetic traps by spurious binding sites impose constraints on TF-DNA interaction

Thermodynamics of TF binding

- Compute probability $p(E)$ of specific binding at a functional site:
 - Idealize problem: Neglect unbound state, 1 factor protein in equilibrium between states, random sequence of length $N \gg 1$ with only one functional site

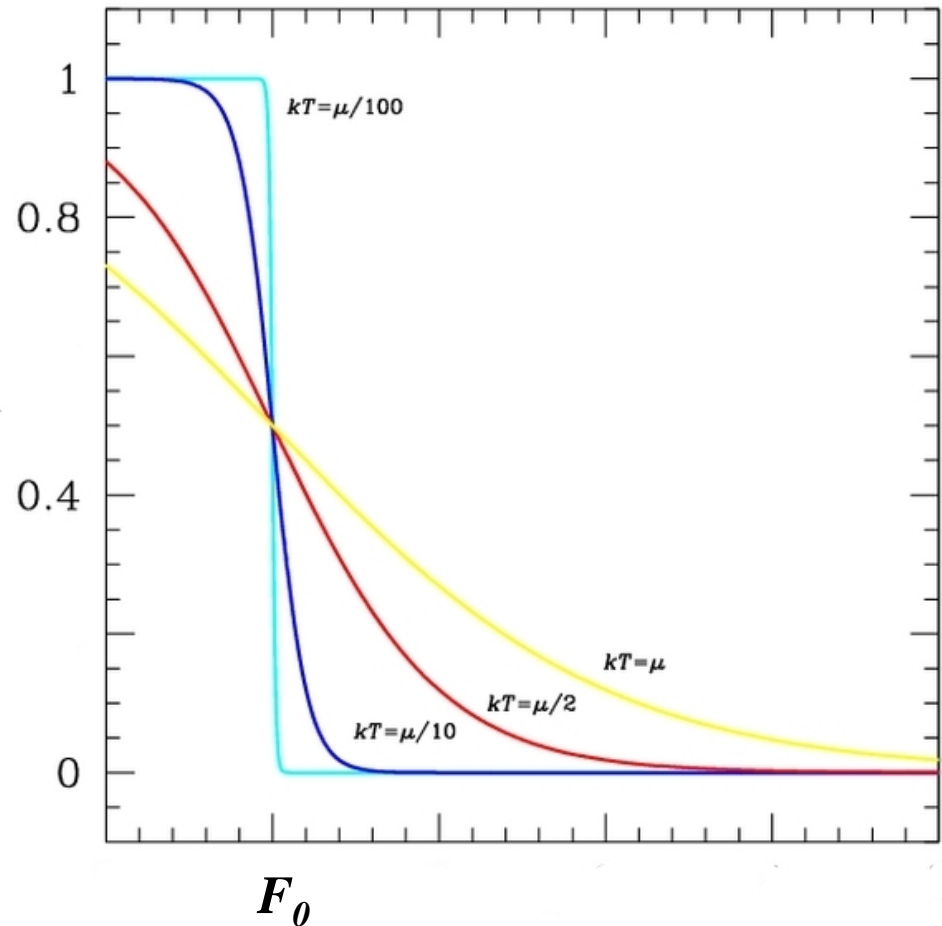
- Use of Boltzmann factors results in

$$\frac{\exp(-\frac{E(r)}{k_B \cdot T})}{\exp(-\frac{E_u}{k_B \cdot T})} \quad p(E) = \frac{1}{1 + \exp(\frac{E - F_0}{k_B \cdot T})}$$

F_0 = free energy of a random sequence

Thermodynamics of TF binding

- Fermi function describes binding probability, with threshold energy $E = F_0$ between strong and weak binding



Thermodynamics of TF binding

- High sensitivity in living cells: single molecules have regulatory effects
- Kinetic traps constrain genomic design
 - Length of TFBS $10^6 \cdot \left(\frac{1}{4}\right)^l \lesssim 1 \Leftrightarrow l \gtrsim \frac{\log 10^6}{\log 4} \approx 10$
 - Binding energy per NT
 - Energy gap between unspecific and optimal binding
- In bacteria, bounds fulfilled as approximate equalities, hence regulation operates just at threshold of single-molecule sensitivity

Implications

- Two parameters allow tuning of regulation
 - Number of TF (time scale of cell cycle)
 - Binding energies (evolutionary time scale)
- Maximal flexibility at single TF sensitivity results in competing design principles
 - Network *programmability* favors larger threshold F_0
 - Stochastic *evolvability* by mutations favors lower threshold F_0

Implications

- Bacteria marginally reach single-molecule sensitivity, which might indicate a compromise between *programmability* and *evolvability*

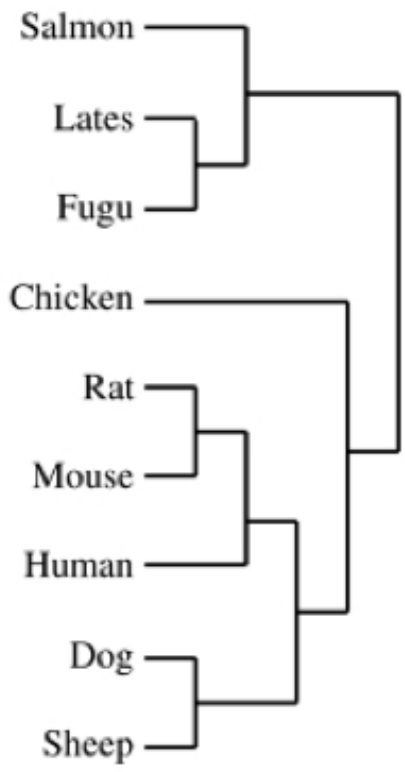
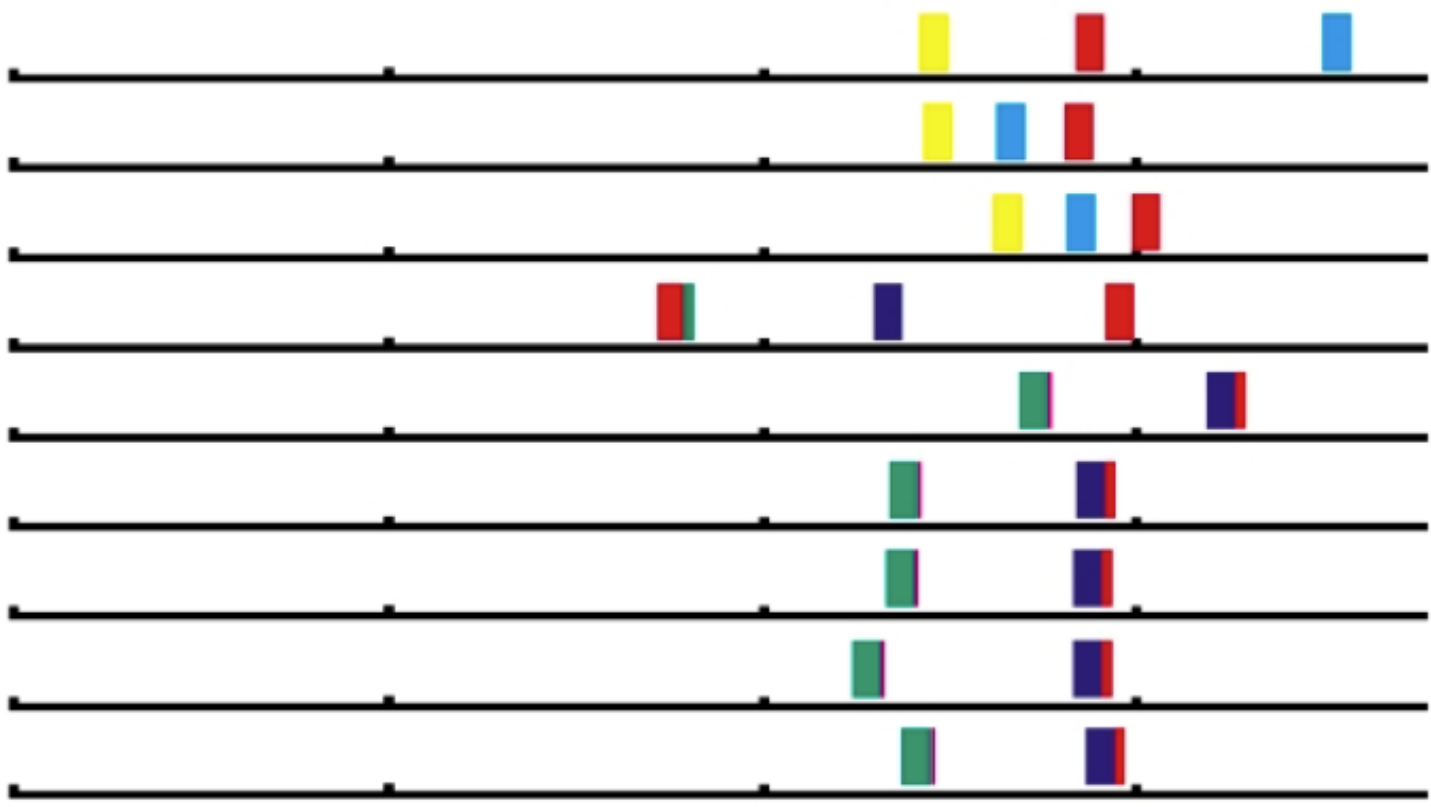
„Binding sites are just complicated enough to work.“

III - Finding Regulatory Elements

- FootPrinter (Blanchette & Tompa, 2003)
- PhyloGibbs (Siddharthan *et al.*, 2005)
- Zhou & Wong 2007
- SAPF (Satija *et al.*, 2008a)
- BigFoot (Satija *et al.*, 2008b)

FootPrinter

- Regulatory elements evolve at slower rate than non-regulatory elements, hence, have higher levels of conservation
- Uses the phylogenetic footprinting method:
 - alignment of homologous regulatory regions
 - multiple species phylogenetic tree
- Doesn't need any known motifs as input:
 - identifies the best conserved motifs between species
 - motifs are used as “*indicators*” of regulatory regions



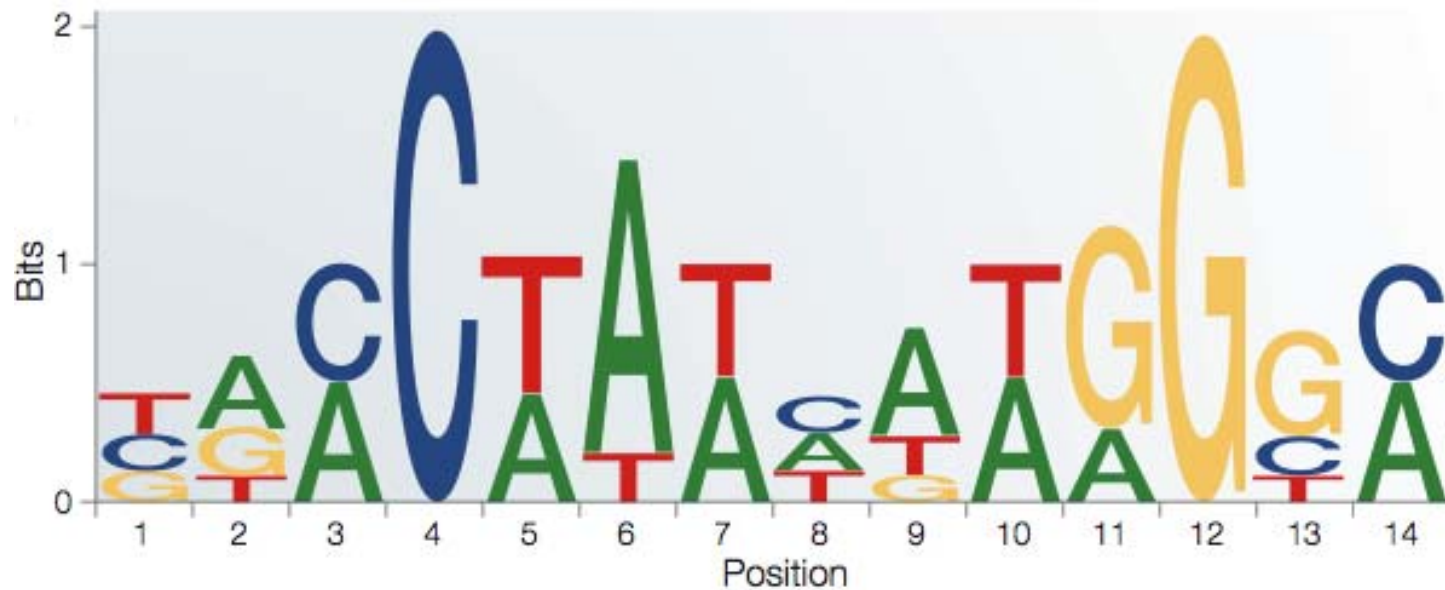
PhyloGibbs

- Enhances FootPrinter by taking non-homologous regions into account:
 - retain patterns of conserved sequence blocks (motifs) and unaligned sequences
 - runs an arbitrary collection of multiple alignments of orthologous intergenic sequences
- Weight matrices can be used to locate putative binding sites.
- For close related species, large sequence blocks can be unambiguously aligned and the search space reduced by pre-aligning them.

Position weight matrix (PWM)

A	-1.93	0.79	0.79	-1.93	0.45	1.50	0.79	0.45	1.07	0.79	0.00	-1.93	-1.93	0.79
C	0.45	-1.93	0.79	1.68	-1.93	-1.93	-1.93	0.45	-1.93	-1.93	-1.93	-1.93	0.00	0.79
G	0.00	0.45	-1.93	-1.93	-1.93	-1.93	-1.93	-1.93	0.66	-1.93	1.30	1.68	1.07	-1.93
T	0.15	0.66	-1.93	-1.93	1.07	0.66	0.79	0.00	0.00	0.79	-1.93	-1.93	-0.66	-1.93

Sequence logo



Zhou & Wong 2007

- Enhances PhyloGibbs motif prediction by using regulatory modules (patterns of TFBS):
 - to identify patterns of motif blocks
 - no fixed optimal alignment, but dynamically updated alignment of orthologous sequences
- Module information captured through coupled Hidden Markov Models (HMM)

SAPF

- Drawback of FootPrinter:
 - uses only one optimizing alignment, hence might miss orthologous segments due to specific alignment
- Similar to PhyloGibbs, enhances FootPrinter by considering statistical alignment:
 - considers many probability weighted alignments using multiple sequence HMM
 - doubling the number of HMM states accounts for phylogenetic footprinting:
 - “fast”, higher levels of divergence as in neutral sequences
 - “slow”, divergence as in purifying selection) accounts for phylogenetic footprinting

BigFoot

- Enhances SAPF by allowing for a larger number of sequences
- Uses a Markov Chain Monte Carlo approach:
 - samples sequence alignments
 - samples locations of slowly evolving regions

IV – Annotation of signals

- Finding methods revisited: Practical issues
 - Homologous vs. Non-homologous annotation
 - The use of additional information
- Limits of comparative genomics methods
 - A simple model to derive bounds on the number of sequences and feature size

Finding methods

- 2 major classes of approaches:
 - Homologous methods
 - Use the information of relatedness (alignment) to prune search space
 - More efficient
 - Non-homologous methods
 - Able to detect movement of binding sites
 - False positives due to increasing noise (background conservation)

Finding methods

- Improve finding methods by use of mRNA expression data
 - Combining phylogenetic footprinting with information of co-regulation (e.g. from microarray profiling, chromatin immunoprecipitation)
 - Relies on availability of such data

A model of statistical power

- Planning comparative genome sequencing
 - How many more genomes are needed to look at smaller conserved features (exons > regulatory sites > single nucleotides)?
 - When is the point of diminishing returns reached?
- Scaling relationship between genome number, evolutionary distance, feature size

A model of statistical power

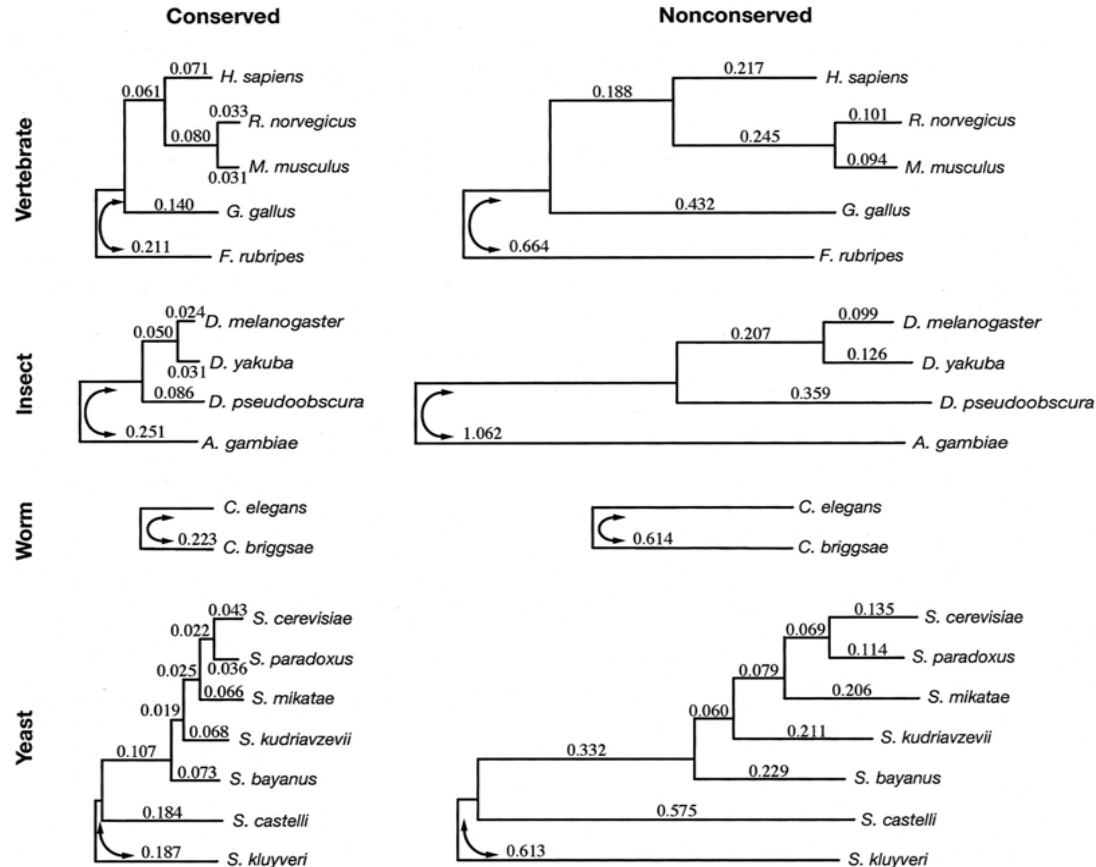
- Lots of assumptions later...
 - For given evolutionary distance, the number of genomes needed for a constant level of statistical stringency scales inversely with the size of the conserved feature
 - For short evolutionary distance, the number of genomes scales inversely with distance

V – Evolution of regulation

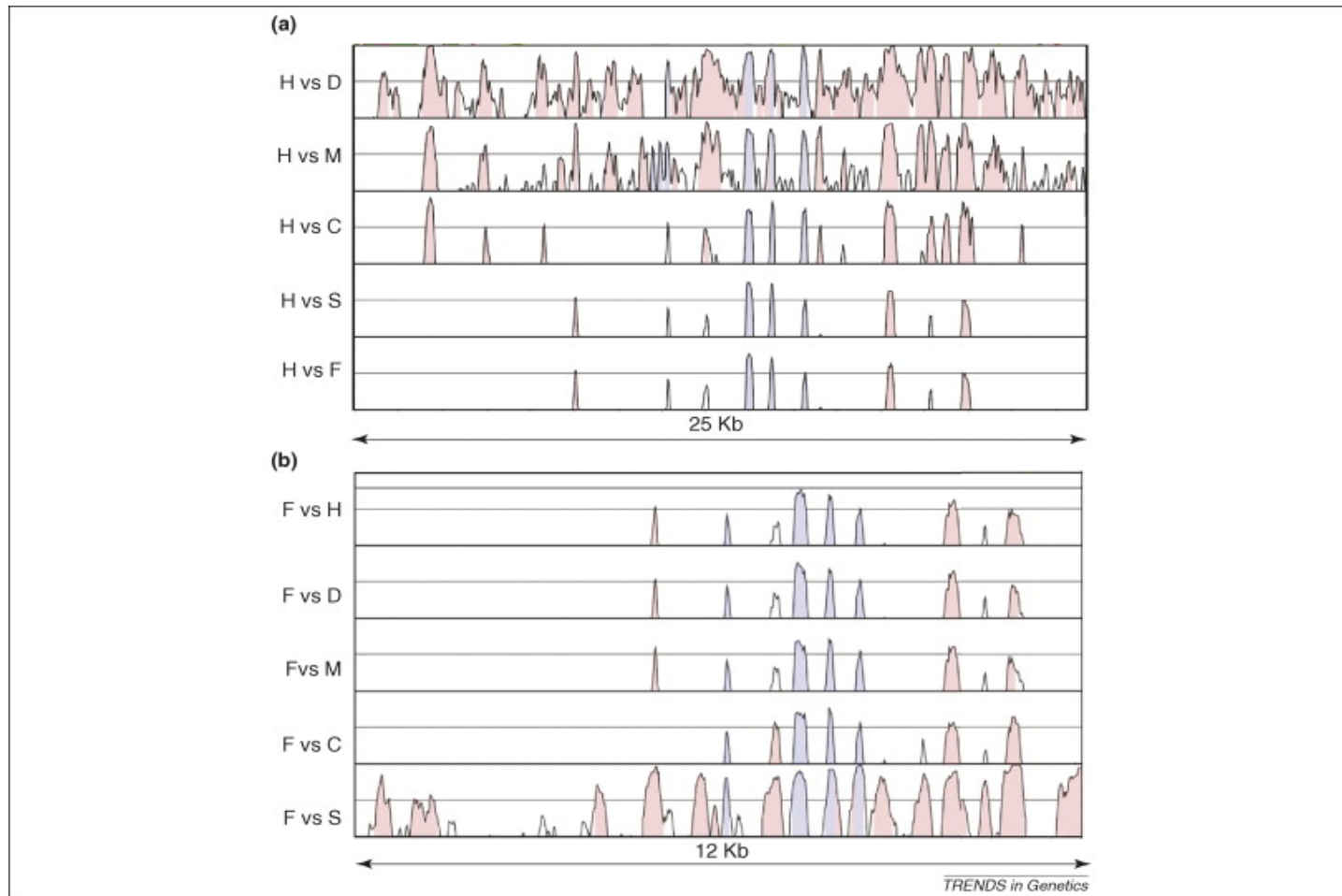
- Regulatory elements
- Summary

Regulatory elements evolution

Understanding the mechanisms of gene regulation, and how evolution of the pattern of gene regulation contributes to morphological and phenotypic differences among organisms are fundamentally important goals in the genome era



Regulatory elements evolution



Conservation is defined by the baseline species. Different views of sequence conservation depending on the species used for comparison. (a) The 5' region of the human (H) Pax7 gene on chromosome is aligned with equivalent regions from dog (D), mouse (M), chicken (C), Fugu (F) and stickleback (S). (b) By contrast, pairwise comparison of sequences with the Fugu region allows the identification of several conserved sequences that are shared between Fugu and stickleback.

Summary

- The understanding of regulatory gene mechanisms has been improved through the analysis of sequence evolution (phylogenetic footprinting) and biophysics of transcription factors and binding sites.

Challenges:

- Need for more biological information about regulatory elements
- Computational analysis limitation (time improving and large number of sequences)
- Evolutionary meaning

“We are drowning in information, while starving for wisdom.”

Edward O. Wilson