BIOZENTRUM

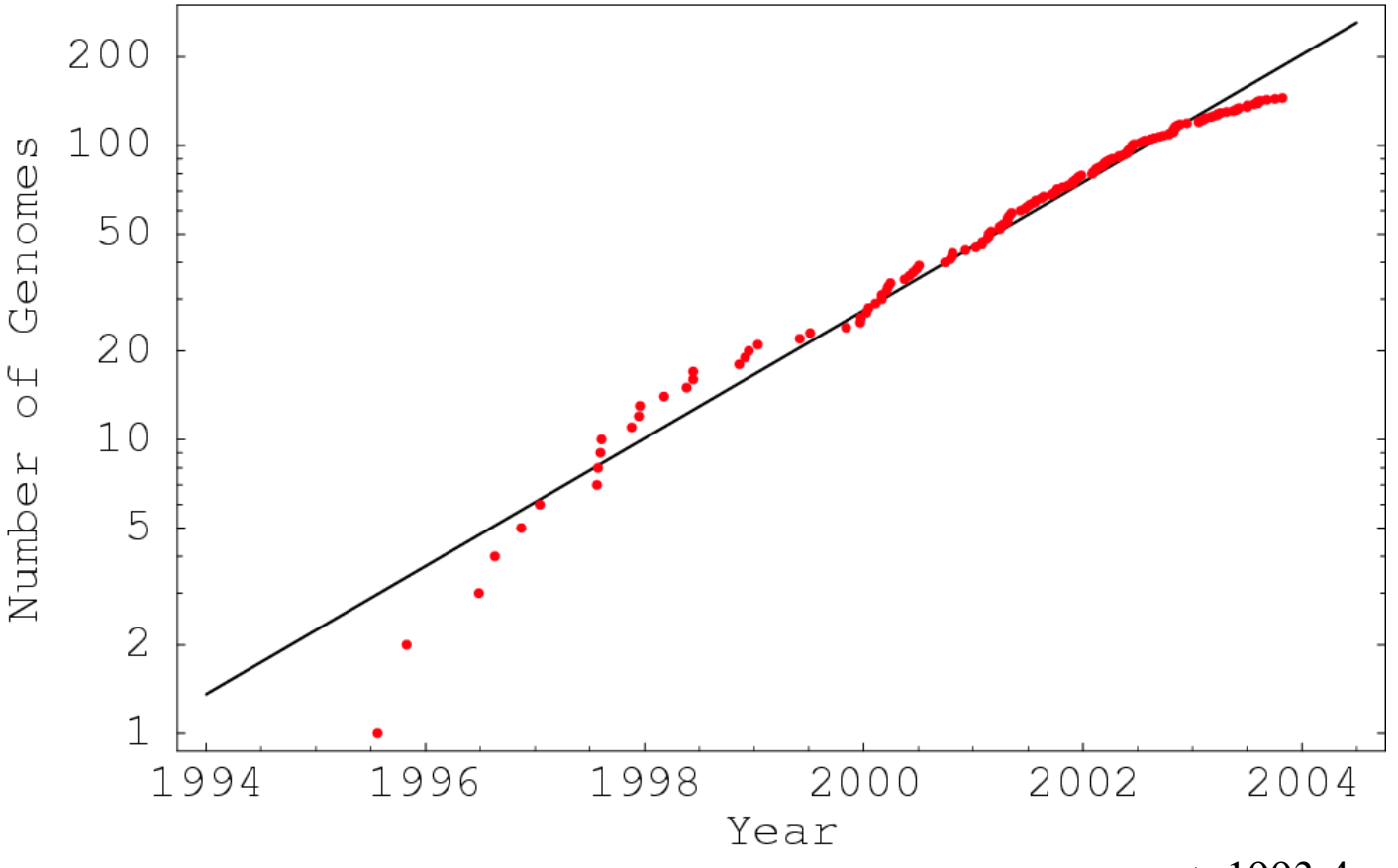
Inferring principles of regulatory design using comparative genomics

Erik van Nimwegen

Division of Bioinformatics
Biozentrum, Universität Basel,
Switzerland

BIOZENTRUM

Exponential growth of the number of sequenced genomes



1 000 genomes by 2007
1 000 000 genomes by 2020

Fit: $N = 2^{\frac{t-1993.4}{1.38}}$

BIOZENTRUM

Topics

1. Genome-wide discovery of new bacterial regulons.

2. Identifying developmental enhancer modules.

3. Scaling in the functional content of genomes.

BIOZENTRUM

Topics

1. Genome-wide discovery of new bacterial regulons.

2. Identifying developmental enhancer modules.

3. Scaling in the functional content of genomes.

BIOZENTRUM

Analyzing genome wide transcription regulation in bacteria

The E. coli transcription regulatory network:

- 210,000 papers on E. coli.
- 18,700 on transcription in E. coli.
- 2500 operons and regulatory regions. 10% have been studied experimentally.
- 300 transcription factors (TFs).
- 400 known TF binding sites (excluding sigma-factor sites) for 54 TFs.
- Expected 3000-6000 sites in total.

We know only (a biased) 10% of the regulatory network.

Outline

1. Find putative binding sites by comparing regulatory regions of orthologous genes in different bacterial genomes.

2. Infer putative *regulons* by determining which binding sites are recognized by the same transcription factor through probabilistic clustering.

BIOZENTRUM

The weight matrix representation of transcription factor binding sites

Alignment of known **fruR** binding sites:

```
AAGCTGAATCGATTTTATGATTTGGT
AGGCTGAATCGTTTCAATTCAGCAAG
CTGCTGAATTGATTCAGGTCAGGCCA
GTGCTGAAACCATTCAAGAGTCAATT
GTGGTGAATCGATACTTTACCGGTTG
CGACTGAAACGCTTCAGCTAGGATAA
TGACTGAAACGTTTTTGCCCTATGAG
TTCTTGAAACGTTTCAGCGCGATCTT
ACGGTGAATCGTTCAAGCAAATATAT
GCACTGAATCGGTTAACTGTCCAGTC
ATCGTTAAGCGATTCAGCACCTTACC
**gcTGAAtCG*TTcAg**c******
```

$$w_\alpha^i = \textbf{Probability of finding base } \alpha \textbf{ at position } i.$$

$$\textbf{For instance: } w_A^3 = 0.267 \quad , w_C^3 = 0.2 \quad , w_G^3 = 0.467 , \; w_T^3 = 0.067$$

Probability that sequence *s* is a binding site for the factor represented by ***w:***

$$P(s \mid w) = \prod_{i=1}^{l} w_{s_i}^{i}$$

Extraction of putative binding sites

1. For each E. coli gene, find orthologs in other bacterial genomes.
2. Find putative binding sites upstream of orthologous genes:
   A. Align regulatory regions.Identify conserved stretches of regulatory region.
      (Rajewsky et al., Genome Res. 2002).
   B. Find multiple local alignment that maximizes total conservation.
      (Gibbs Sampler, McCue et al. Nucleic Acids Res. 2001).

Extraction of putative binding sites

1. For each E. coli gene, find orthologs in other bacterial genomes.
2. Find putative binding sites upstream of orthologous genes:
   A. Align regulatory regions.Identify conserved stretches of regulatory region.
      (Rajewsky et al., Genome Res. 2002).
   B. Find multiple local alignment that maximizes total conservation.
      (Gibbs Sampler, McCue et al. Nucleic Acids Res. 2001).

ATTTAGACGTCTGGAT
ATTTAGACGTCTGGAT
ACTTAGCCGTCTAGAC
ATCTAGACATCTAGAA
TTCTAGACGTCTAGAA
ATTTAGACGTCTAGAC

Maximize “total conservation”.

A few thousand groups of putative sites in the E. coli genome.

BIOZENTRUM

Clustering putative binding sites

If *S* is a set of binding sites and *w* a weightmatrix then the probability that all these sites came from a single weightmatrix is given by:

$$P(S) = \int \mathbf{dw} \prod_{\mathbf{s} \in \mathbf{S}} \mathbf{P(s \mid w)} = \prod_{\mathbf{i=1}}^{\mathbf{l}} \frac{\mathbf{3!} n_A^i! n_C^i! n_G^i! n_T^i!}{(n+3)!}$$

If *C* describes a partition of the data *D* into clusters *Sc* then the probability of the partition *C* is

$$P(C \mid D) = \prod_c P(Sc) / Z$$

We assign a probability to each possible way of dividing the sites into clusters.

BIOZENTRUM

Example

Assume the data set consists of only these four sequences:

s1 = aaacgattcagttaggc
s2 = tcaagctaggtattacc
s3 = aaccgtttcgattcgga
s4 = tcgagaaaggtatcagc

The probabilities *P(C|D)* for several partitions (ways of clustering) the data:

P(s1,s3) P(s2,s4) = 0.543
aaacgattcagttaggc
aaccgtttcgattcgga
tcaagctaggtattacc
tcgagaaaggtatcagc

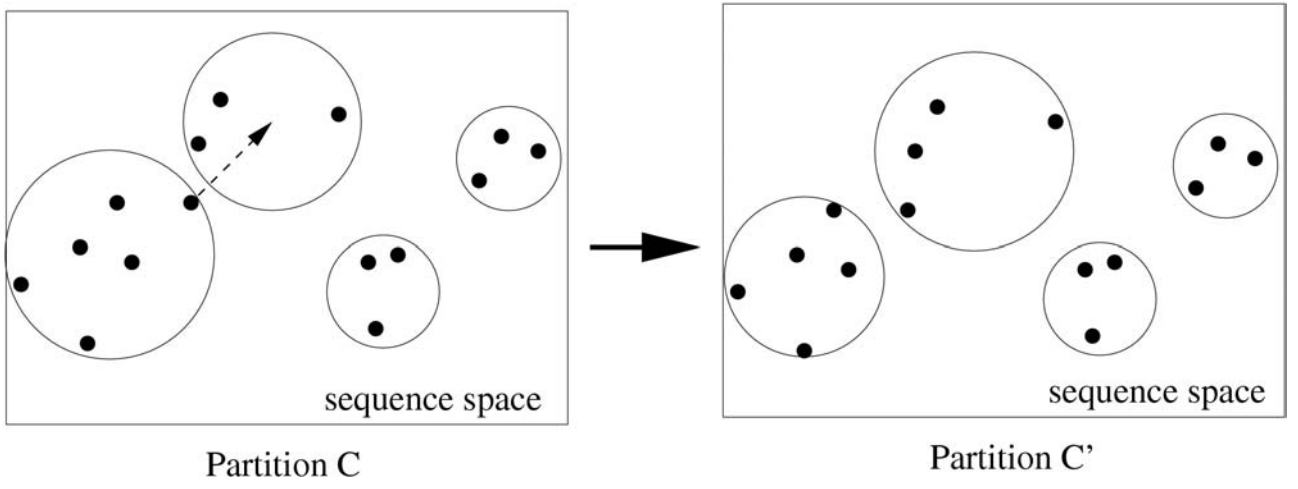
P(s1,s3,s4) P(s2) = 0.185
aaacgattcagttaggc
aaccgtttcgattcgga
tcgagaaaggtatcagc
tcaagctaggtattacc

P(s1,s2,s4) P(s3) = 0.084
aaacgattcagttaggc
tcaagctaggtattacc
tcgagaaaggtatcagc
aaccgtttcgattcgga

P(s1,s3) P(s2) P(s4) = 0.086
aaacgattcagttaggc
aaccgtttcgattcgga
tcaagctaggtattacc
tcgagaaaggtatcagc

P(s1)P(s3) P(s2,s4) = 0.084
aaacgattcagttaggc
aaccgtttcgattcgga
tcaagctaggtattacc
tcgagaaaggtatcagc

Sampling the probability distribution.



Standard Monte-Carlo Markov chain moves:

1. Propose random move from *C* to *C'.*
2. Accept when *P(C)>P(C')* or with *P(C)/P(C')* when *P(C)<P(C')* .

Identify "stable" clusters of sites that consistently "stick together" during sampling.

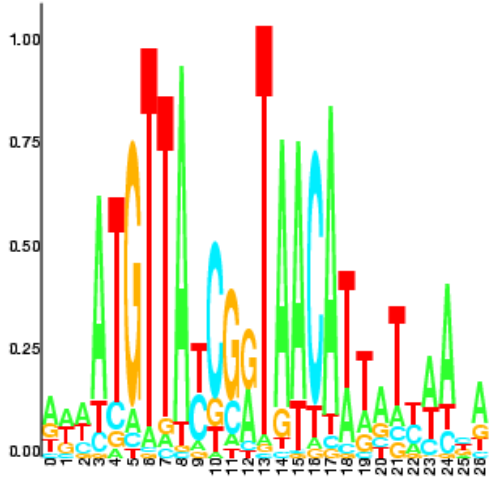
Example Cluster

Score	Distance to operon	Sense	Operon structure
0.981	110	+	idnK b4268
	107	-	idnD b4267 .. 24 .. idnO b4266 .. 62 .. idnT b4265 .. 67 .. idnR b4264 .. 78 .. yjgR b4263
0.981	90	-	gntK b3437 .. 4 .. gntU_1 b3436 .. 6 .. gntU_2 b3435
0.981	25	+	b2740 b2740
0.981	169	+	gntT b3415
0.981	28	+	gntT b3415
0.917	81	+	idnK b4268
	136	-	idnD b4267 .. 24 .. idnO b4266 .. 62 .. idnT b4265 .. 67 .. idnR b4264 .. 78 .. yjgR b4263
0.179	386	+	yajF b0394 .. -21 .. b0395 b0395
0.179	14	+	yajF b0394 .. -21 .. b0395 b0395
0.072	38	+	yegT b2098 .. -3 .. b2099 b2099 .. -3 .. b2100 b2100
	147	-	b2097 b2097

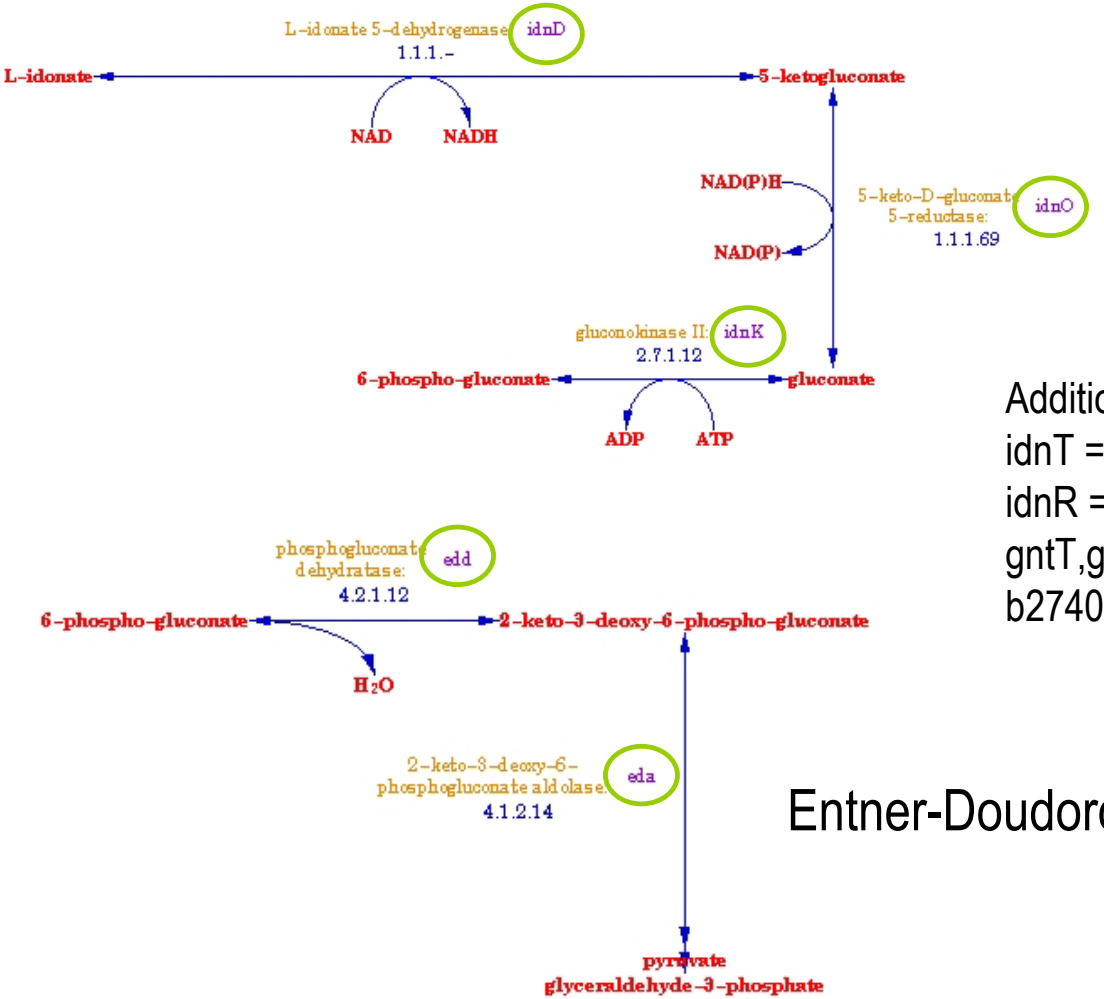
Probability that site belongs to cluster

Genes downstream of site.

Inferred Weight Matrix
(sequence motif recognized
by transcription factor.)



GntII system: catabolism of L-idonate with D-gluconate as an intermediate



Bausch et. al.
J. Bacteriol. Vol 180,
p. 3074 (1998)

Additional genes in cluster:
idnT = L-idonate transporter
idnR = regulator idnDOTR operon
gntT,gntU = gluconate transporters.
b2740 = homologue of gluconate permease

Entner-Doudoroff pathway

BIOZENTRUM

Results

van Nimwegen et al. PNAS **99:**7323 (2002)

Clustering 2000 mini-WMs from McCue et al.:
115 stable clusters.
21 correspond to known regulons, 94 new putative regulons.

Clustering 2000 mini-WMs from Rajewsky et al. :
65 stable clusters.
25 correspond to known regulons, 40 new putative regulons.

Roughly 150 new sites for known regulons, and 500 sites for unknown regulons.

http://www.physics.rockefeller.edu/regulons

BIOZENTRUM

Topics

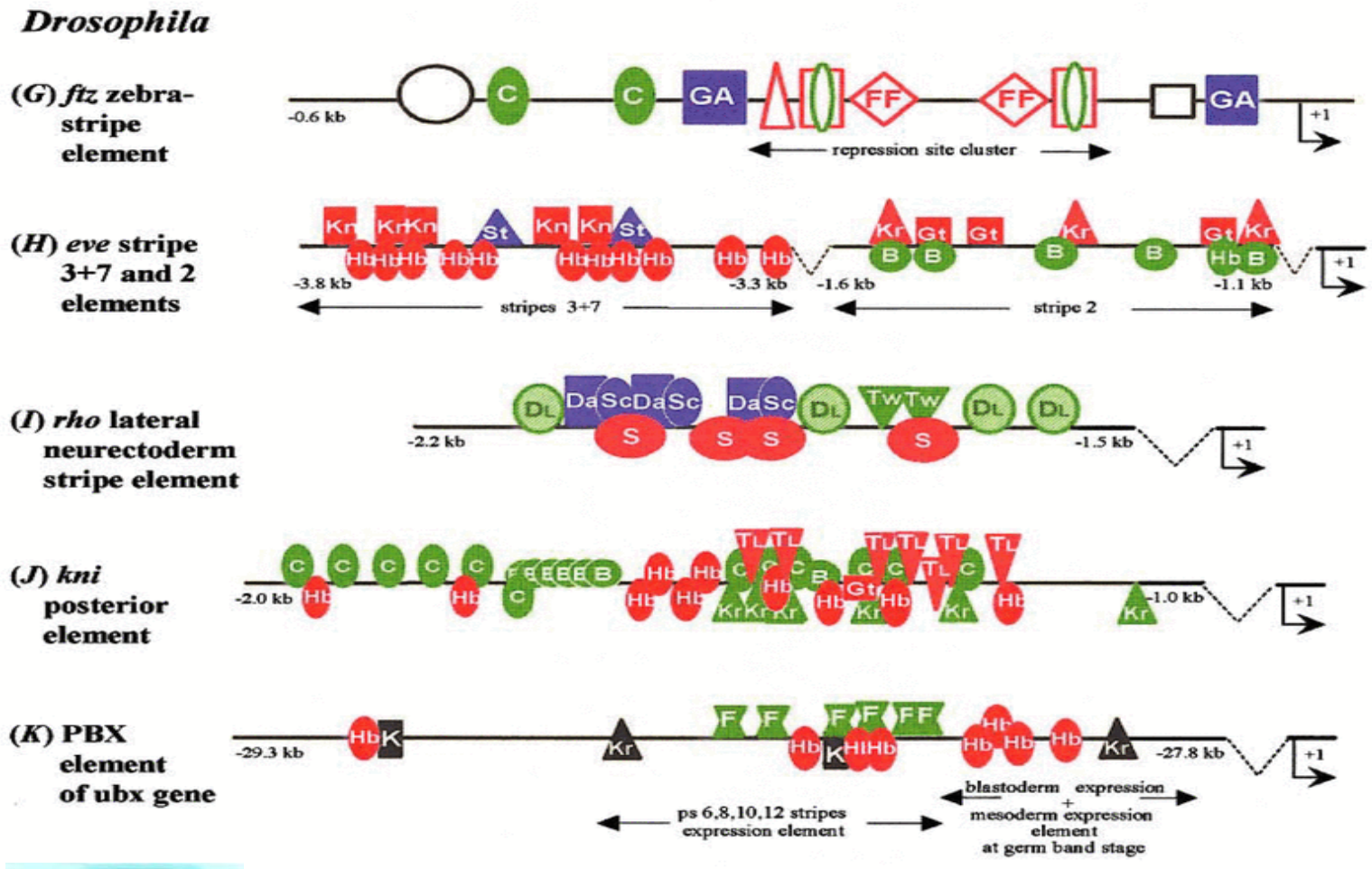
1. Genome-wide discovery of new bacterial regulons.

2. Identifying developmental enhancer modules.

3. Scaling in the functional content of genomes.

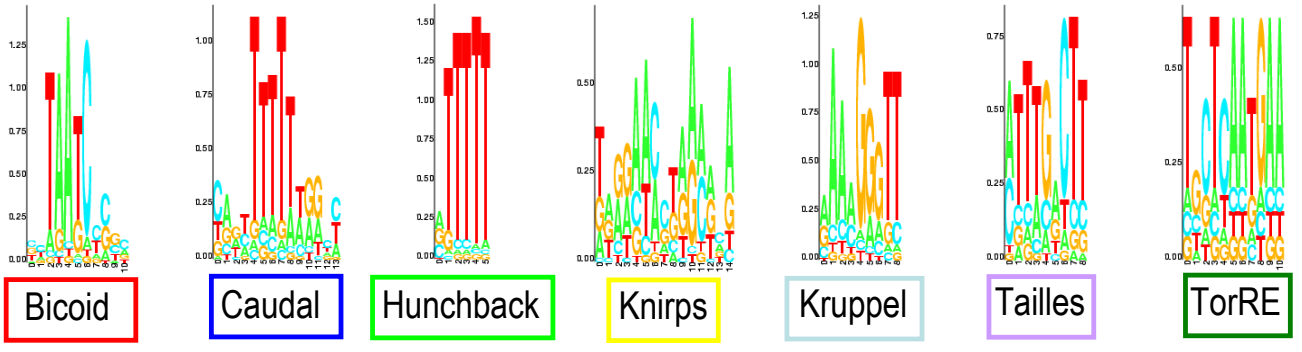
BIOZENTRUM

Regulatory modules in Drosophila

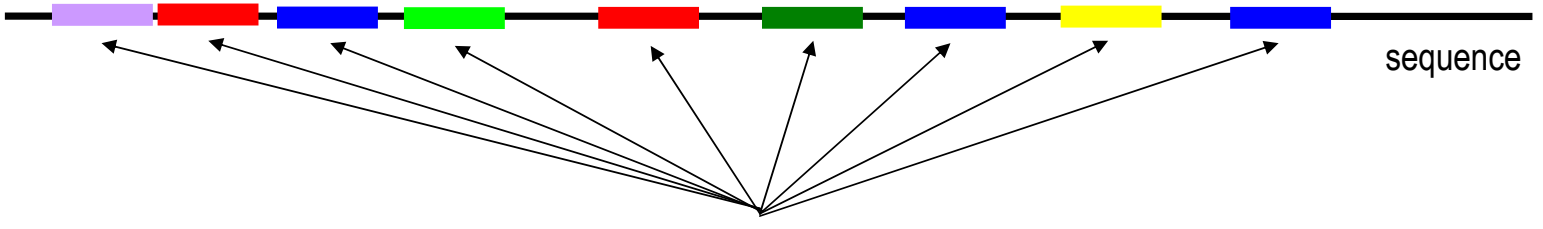


(from Arnone, M. I. and Davidson, E. H., *Development,* **124**(10):1851-64, 1997.)

Parsing a sequence in terms of binding sites



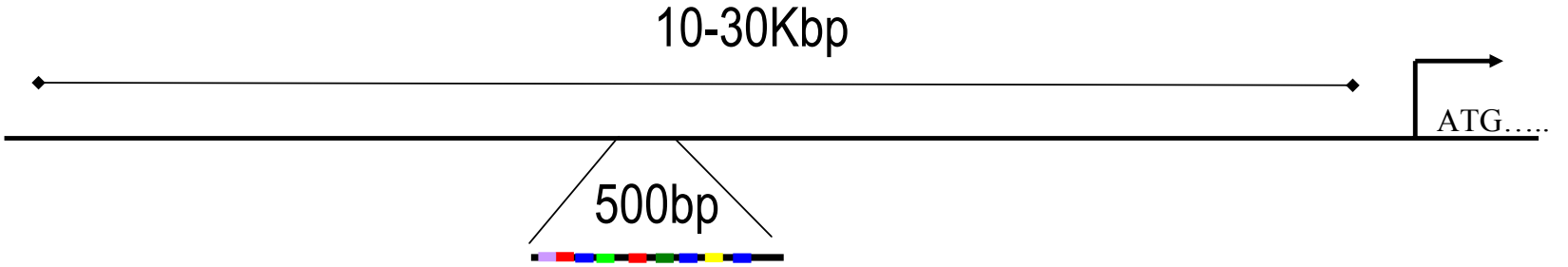
Set of Weight Matrices for the gap gene transcription factors



A `parse' ρ of the sequence *S* in terms of hypothesized binding sites.

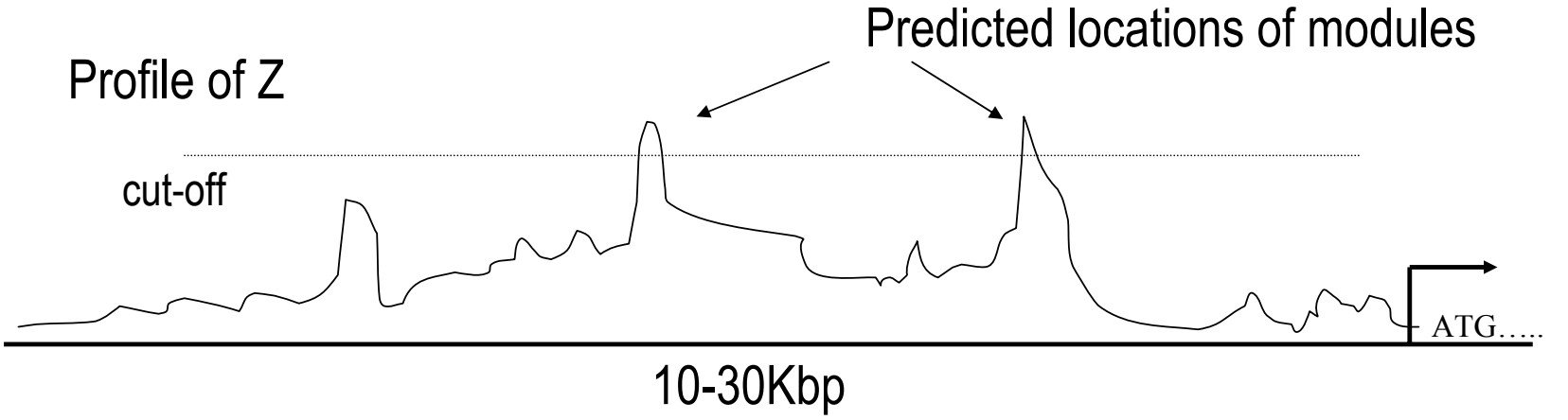
$P(S \mid \rho)$ Probability of the observed sequence given the parse.

Scanning upstream regions for clusters of binding sites



For each 500bp region, calculate the sum of probabilities of all parses *Z*.

$$Z = \sum_{\rho} P(S \mid \rho)$$



Results on gap gene upstream regions

Table 1. Performance of *Stubb* (hcHMM) on gap gene upstream regions. The last column measures the fractional overlap between the known and predicted modules

Gene	Predicted Modules	Score	Known Module	Overlap
eve	2780–3279	27.9	2763–3273	0.98
	5100–5600	17.0	4974–5644	1.00
gt	7360–7859	16.0	7242–8184	1.00
hairy	1340–1839	15.7	829–1760	0.84
	2600–3099	32.7	2601–3147	1.00
	5640–6139	12.3	5831–6132	1.00
	7100–7599	18.6	6396–7551	1.00
kni	4140–4639	15.4	not known	—
	6900–7399	23.2	6926–6992	1.00
	7380–7879	28.7	7422–8998	0.91
Kr	5640–6139	18.2	5668–6389	0.94
run	60–559	15.2	37–862	1.00
	6540–7039	17.3	not known	—
tll	7140–7639	23.9	6997–7476	0.67
	8420–8919	19.6	8564–8946	1.00
	9400–9899	13.7	9418–9592	1.00
hb	2420–2919	16.6	2335–3357	1.00
	9000–9499	14.0	8834–9554	1.00

S. Sinha, E. van Nimwegen and E. Siggia. *Bioinformatics* 19: i292-i301

All known modules are recovered!

Two additional modules predicted (currently being tested).

BIOZENTRUM

Topics

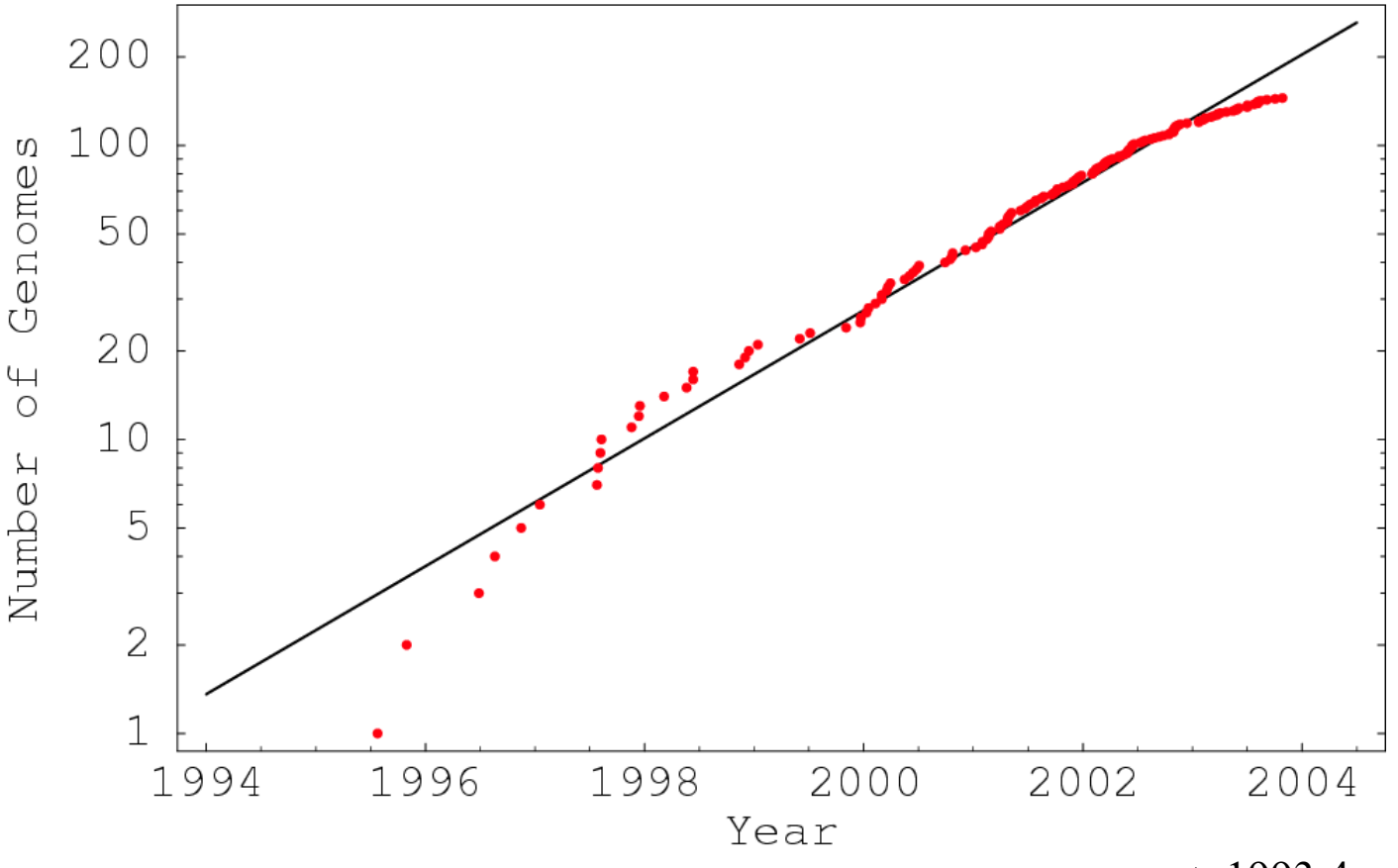
1. Genome-wide discovery of new bacterial regulons.

2. Identifying developmental enhancer modules.

3. Scaling in the functional content of genomes.

BIOZENTRUM

Exponential growth of the number of sequenced genomes



1 000 genomes by 2007
1 000 000 genomes by 2020

Fit: $N = 2^{\frac{t-1993.4}{1.38}}$

BIOZENTRUM

Statistically Comparing Functional Gene Content

1. Define functional categories for gene annotations.

genes involved in metabolism
genes involved in the cell cycle
genes involved in signal transduction
genes involved in transcription regulation

2. Collect all sequenced genomes and run Interpro to annotate genes for known protein domains/families.

EMBL NUCLEOTIDE SEQUENCE DATABASE

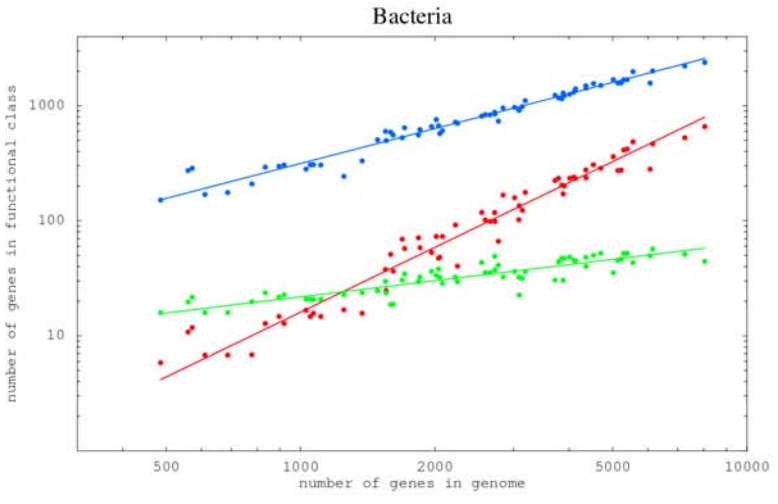
InterPro

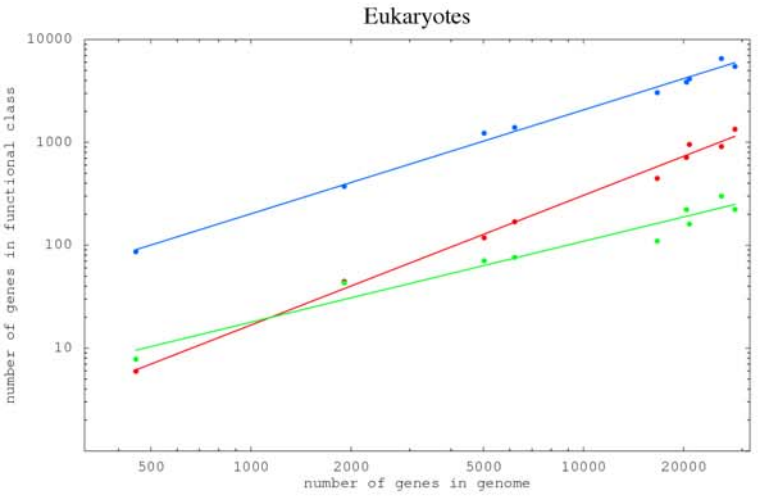
EBI

http://www.ebi.ac.uk/proteome

3. Map Interpro annotation to GO annotation to count number of genes in each GO category.

Example Results



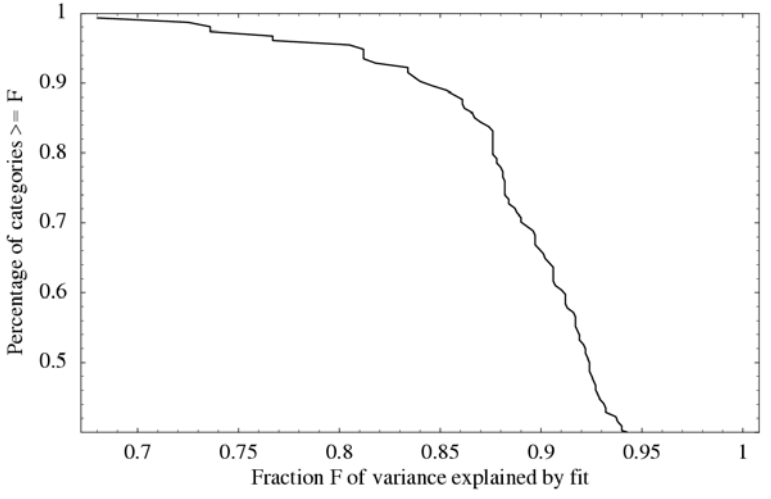


$$n_c = A_c n^{\alpha_c} \Leftrightarrow \log(n_c) = C_c + \alpha_c \log(n)$$

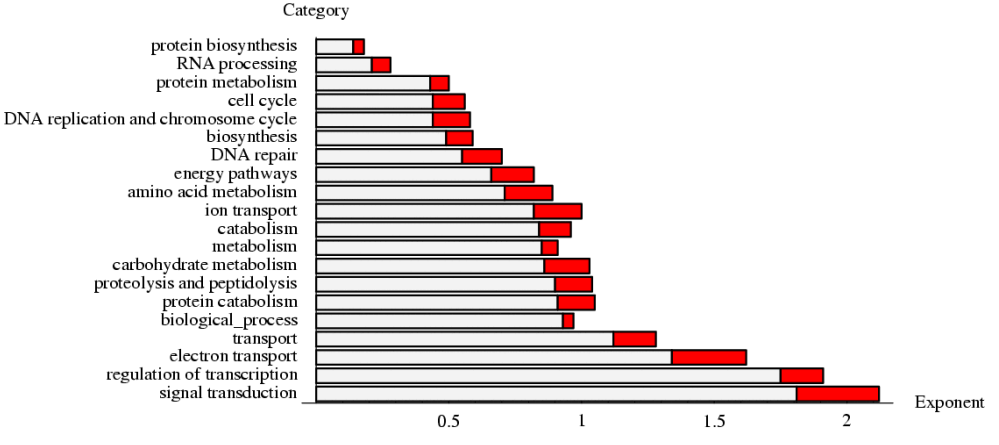
Exponent α_c = Slope of the line for category *c*.

Exponents	Bacteria	Eukaryotes
Metabolism	1.0 +/- 0.1	1.0 +/- 0.2
Transcription regulation	1.9 +/- 0.14	1.3 +/- 0.2
Cell cycle	0.5 +/- 0.1	0.8 +/- 0.4

Overview of the Results in Bacteria



Distribution of the quality of the power-law fit for all 154 categories with at least one match in each genome.



Examples of the observed exponents α_c

BIOZENTRUM

Evolutionary model

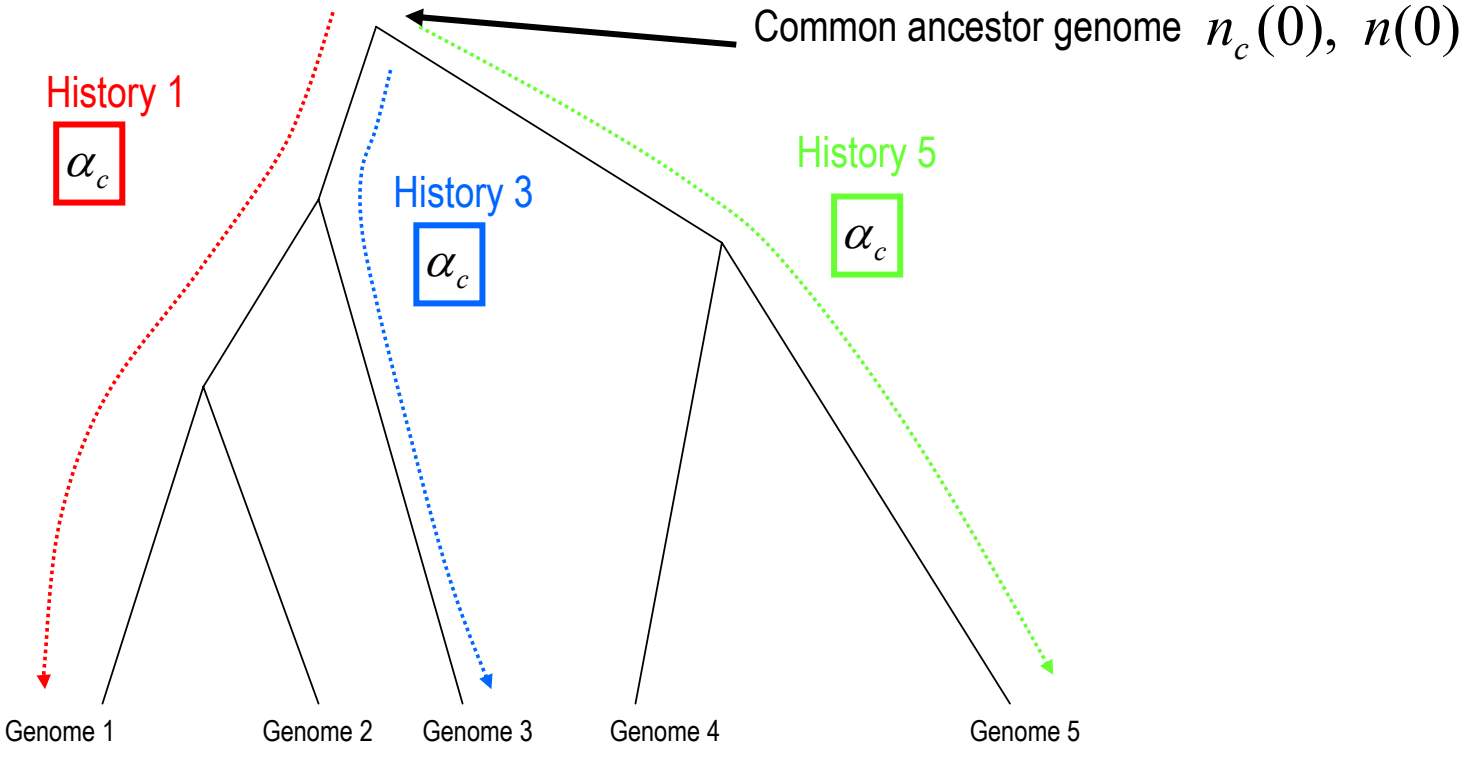
- Consider the evolutionary history of a single genome.
- Think of the genome as a population of genes that are reproducing (duplication) and dying (deletion).

$n_c(t)$	The number of genes in category *c* at time *t*.
$n(t)$	The total number of genes in the genome at time *t*.
$\lambda_c(t)$	The rate of duplication of genes in category *c* at time *t*.
$\mu_c(t)$	The rate of deletion of genes in category *c* at time *t*.
$\lambda(t)$	The total rate of duplication in the genome at time *t*.
$\mu(t)$	The total rate of deletion in the genome at time *t*.

Then one has at any point in time:

$$\frac{n_c(t)}{n_c(0)} = \left(\frac{n(t)}{n(0)}\right)^{\alpha_c}, \quad \alpha_c = \frac{\langle \lambda_c - \mu_c \rangle_{\text{genome history}}}{\langle \lambda - \mu \rangle_{\text{genome history}}}$$

Evolutionary histories of multiple genomes



For each genome separately holds:

$$\frac{n_c(t)}{n_c(0)} = \left(\frac{n(t)}{n(0)}\right)^{\alpha_c}$$

In order for all to fall on the **same** line we need:

α_c = α_c = α_c

The exponents correspond to evolutionary constants.

BIOZENTRUM

Acknowledgments

Collaborators (these projects)**:**

- Nikolaus Rajewsky (*NYU, New York)*
- Eric Siggia (*the Rockefeller University, New York*)
- Saurabh Sinha (*the Rockefeller University, New York*)
- Mihaela Zavolan (*Biozentrum, Basel*)

References:

1. Probabilistic clustering of sequences: Inferring new bacterial regulons by comparative genomics. E. van Nimwegen, M. Zavolan, N. Rajewsky, E.D. Siggia *Proc. Acad. Natl. Sci. USA* **99** 7323-7328 (2002).
2. A probabilistic method to detect regulatory modules. S. Sinha, E. van Nimwegen, E. D. Siggia *Bioinformatics* **19** i292-i301 (2003).
3. Scaling laws in the functional content of genomes. E. van Nimwegen *Trends in Genetics* **19** 479-484 (2004).