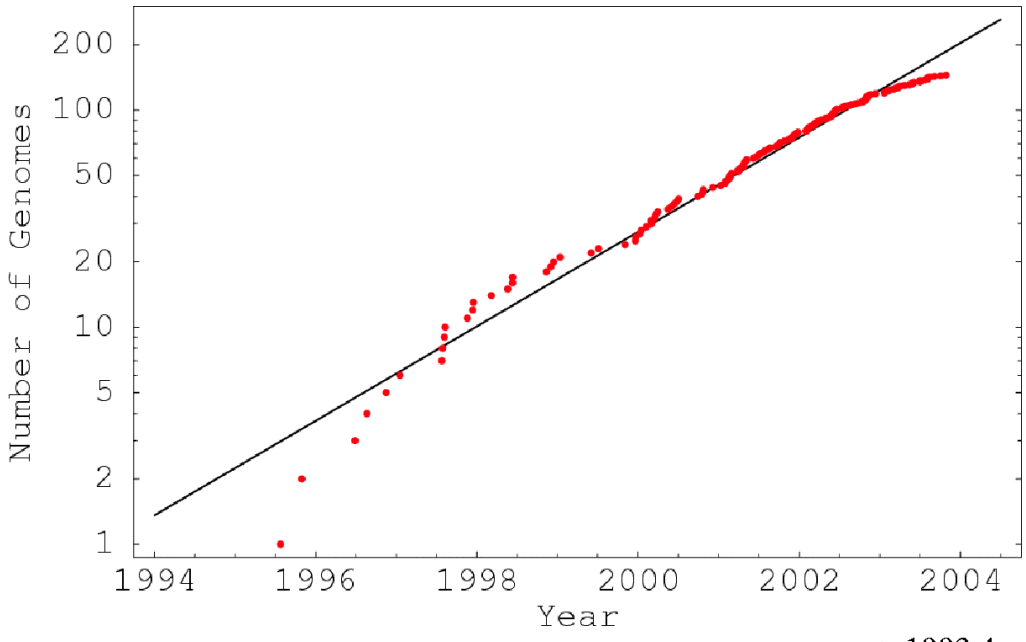
Scaling laws in the functional content of genomes

Erik van Nimwegen
Division of Bioinformatics
Biozentrum, Universität Basel,
Swiss Institute of Bioinformatics

Exponential growth of the number of sequenced genomes



1 000 genomes by 2007
1 000 000 genomes by 2020

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

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 count the number of genes in each of the functional categories.

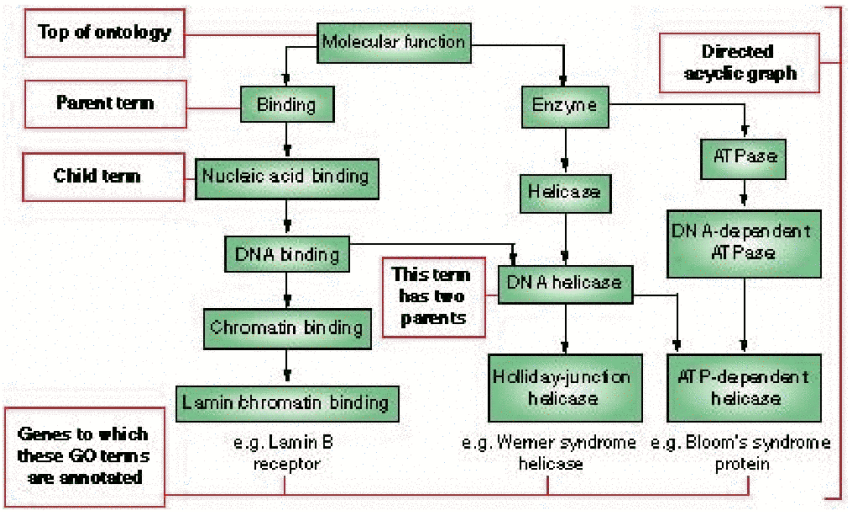
3. Perform quantitative analysis of the distribution of this `functional gene content' across genomes.

Defining Functional Categories:

The Gene Ontology

http://www.geneontology.org

- A collaborative effort between major genome repositories.
- Defines a hierarchy of functional annotations of genes.
- Three hierarchies: molecular function, biological process, and cellular component.



Assigning genes to GO categories:

The Interpro protein domain and family database

InterPro Databases

InterPro Member Databases	
swissprot	The SWISS-PROT database consists of sequence entries. It contains high-quality annotation, is non-redundant and cross-referenced to many other databases.SWISS-PROT is accompanied by TrEMBL, a computer-annotated supplement to SWISS-PROT. TrEMBL contains the translations of all coding sequences (CDS) present in the EMBL Nucleotide Sequence Database, which are not yet integrated into SWISS-PROT.
proSite	PROSITE is a database of protein families and domains. It consists of biologically significant sites, patterns and profiles that help to reliably identify to which known protein family (if any) a new sequence belongs.
Pfam	Pfam is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains.
PRINTS	PRINTS is a compendium of protein fingerprints. A fingerprint is a group of conserved motifs used to characterise a protein family; its diagnostic power is refined by iterative scanning of a composite of SWISS-PROT + SP-TrEMBL. Usually the motifs do not overlap, but are separated along a sequence, though they may be contiguous in 3D-space. Fingerprints can encode protein folds and functionalities more flexibly and powerfully than can single motifs, their full diagnostic potency deriving from the mutual context afforded by motif neighbours.
ProDom	The ProDom protein domain database consists of an automatic compilation of homologous domains. Current versions of ProDom are built using a novel procedure based on recursive PSI-BLAST searches (Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W & Lipman DJ, 1997, Nucleic Acids Res., 25:3389-3402; Gouzy J., Corpet F. & Kahn D., 1999, Computers and Chemistry 23:333-340.) Large families are much better processed with this new procedure than with the former DOMAINER program (Sonnhammer, E.L.L. & Kahn, D., 1994, Protein Sci., 3:482-492).
SMART	SMART (a Simple Modular Architecture Research Tool) allows the identification and annotation of genetically mobile domains and the analysis of domain architectures. More than 500 domain families found in signalling, extracellular and chromatin-associated proteins are detectable. These domains are extensively annotated with respect to phyletic distributions, functional class, tertiary structures and functionally important residues. Each domain found in a non-redundant protein database as well as search parameters and taxonomic information are stored in a relational database system. User interfaces to this database allow searches for proteins containing specific combinations of domains in defined taxa.
TIGR tigr fams	TIGRFAMs is a collection of protein families, featuring curated multiple sequence alignments, Hidden Markov Models (HMMs) and annotation, which provides a tool for identifying functionally related proteins based on sequence homology. Those entries which are "equivalogs" group homologous proteins which are conserved with respect to function.
PIR SuperFamily	PIR Superfamily (PIRSF) is a classification system based on evolutionary relationship of whole proteins. Members of a superfamily are monophyletic (evolved from a common evolutionary ancestor) and homeomorphic (homologous over the full-length sequence and sharing a common domain architecture). A protein may be assigned to one and only one superfamily. Curated superfamilies contain functional information, domain information, bibliography, and cross-references to other databases, as well as full-length and domain HMMs, multiple sequence alignments, and phylogenetic tree of seed members. PIRSF can be used for functional annotation of protein sequences.
Superfamily	SUPERFAMILY is a library of profile hidden Markov models that represent all proteins of known structure. The library is based on the SCOP classification of proteins: each model corresponds to a SCOP domain and aims to represent the entire SCOP superfamily that the domain belongs to. SUPERFAMILY has been used to carry out structural assignments to all completely sequenced genomes. The results and analysis are available from the SUPERFAMILY website.

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

"Interpro is a database of protein families, domains, and functional sites, in which identifiable features found in known proteins can be applied to unknown protein sequences."

Summary of the functional annotation pipeline

For each genome:

1. Obtain the genome sequence.
2. Find the genes and translate into protein sequence.

For each gene:

1. Run Interpro on the protein sequence.

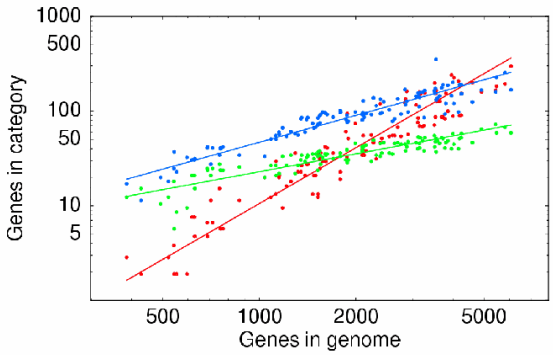
http://www.ebi.ac.uk/integr8

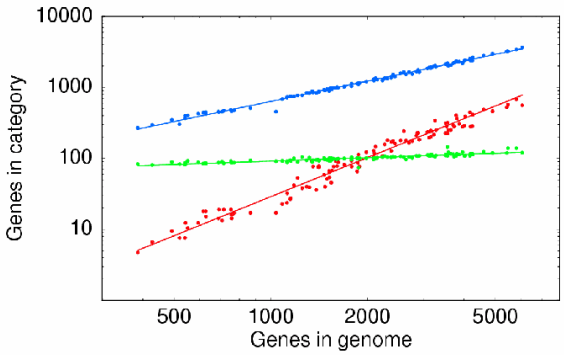
2. Add +1 to each GO-category `hit' by this gene. This includes all parent categories in the hierarchy.
3. Add +1 to number of annotated genes if the gene has one or more Interpro hits.

Question:

How does the number of genes in different categories depend on the total number of genes in the genome?

Example Results in Bacteria



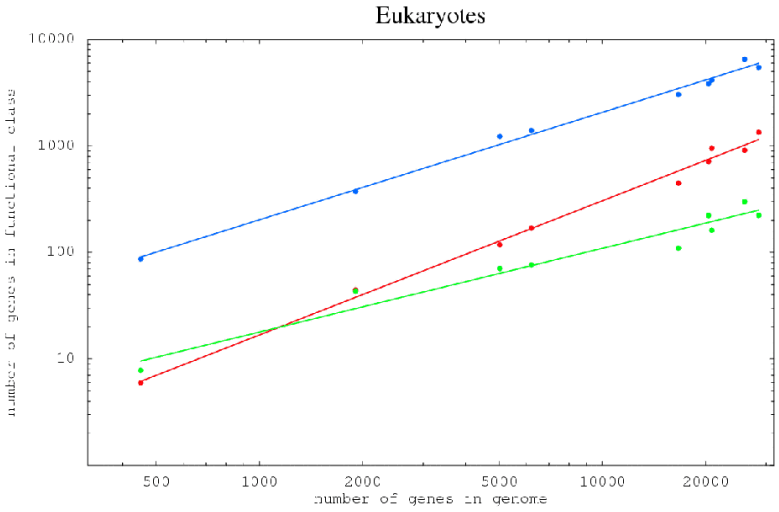


$$n_c = A_c n^{\alpha_c} \Leftrightarrow \log(n_c) = C_c + \alpha_c \log(n)$$

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

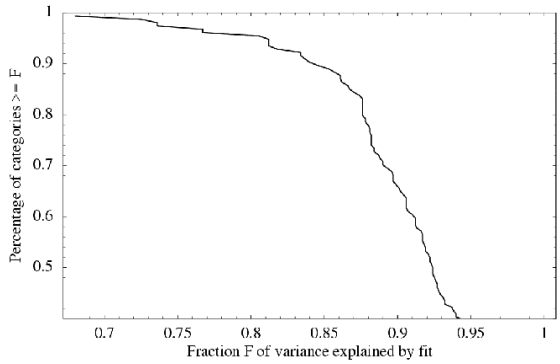
Left panel		Right Panel	
Signal transduction	1.95 +/-0.15	Transcription regulation	1.83+/-0.12
Carbohydrate metabolism	0.94+/-0.1	Biological process	0.94+/-0.04
DNA repair	0.63+/-0.07	Protein biosynthesis	0.15+/-0.03

What about Eukaryotes?

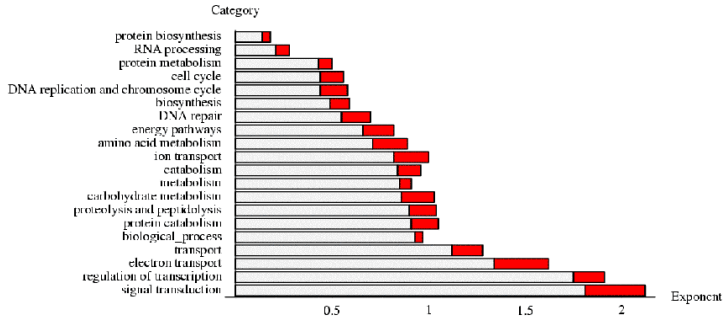


Transcription regulation	1.3 +/- 0.2
Metabolism	1.0 +/- 0.2
Cell cycle	0.8 +/- 0.4

Overview of the Results in Bacteria



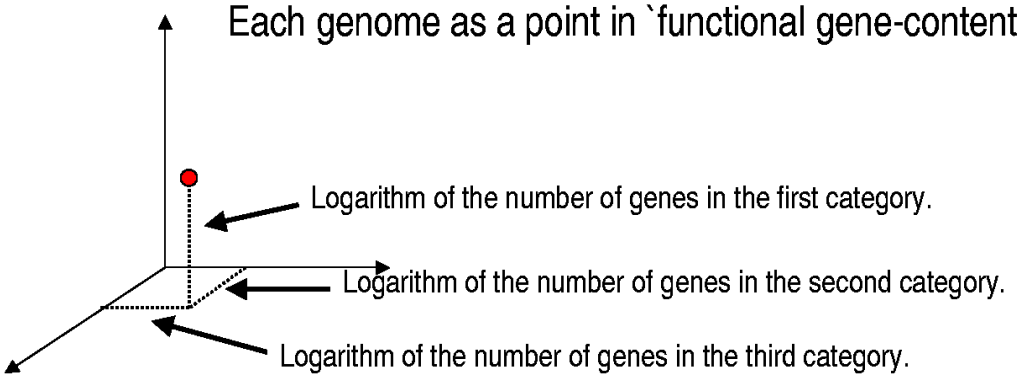
Distribution of the quality of the power-law fit for all 154 categories with at least one match in each genome.



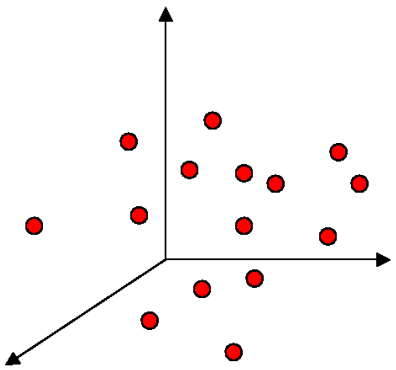
Examples of the observed exponents α_c

Another way of representing the results

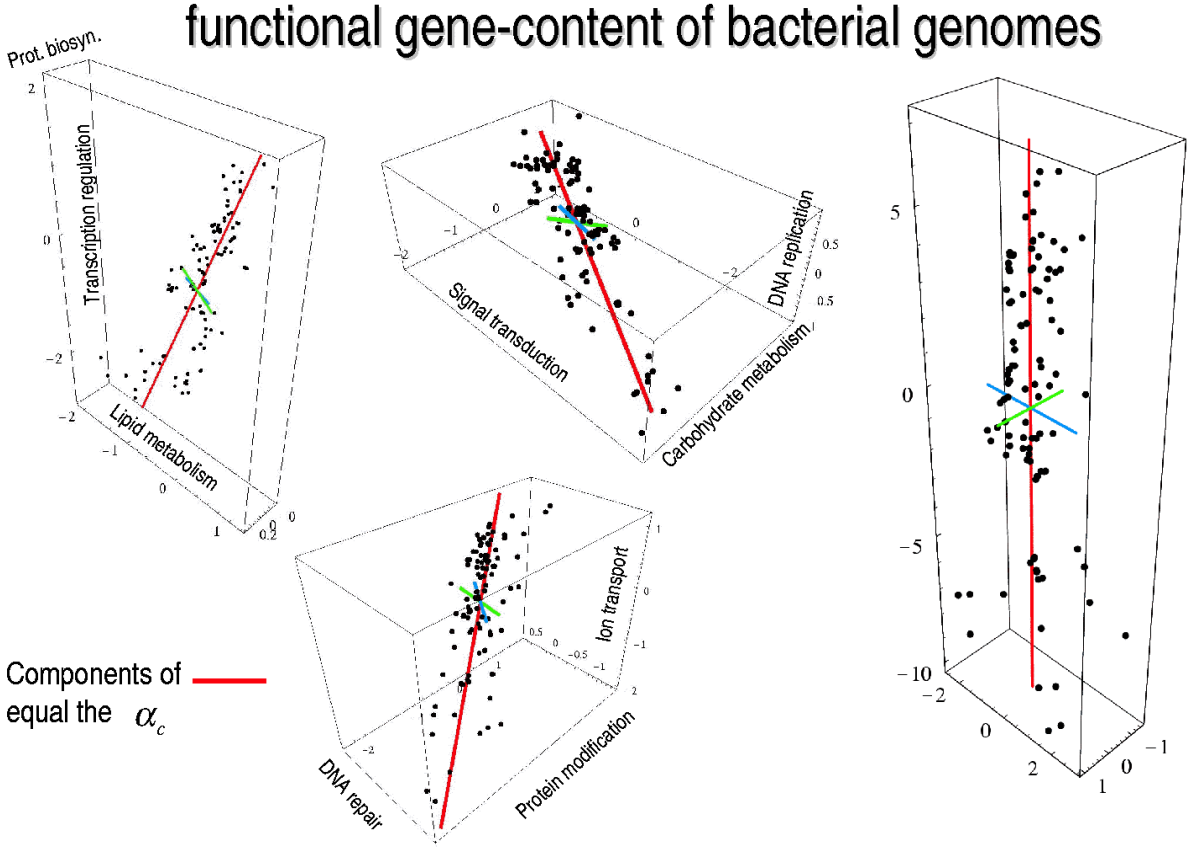
Each genome as a point in `functional gene-content space'.



Genomes form a `cloud' in the functional gene-content space.



The first 3 principal components of the functional gene-content of bacterial genomes

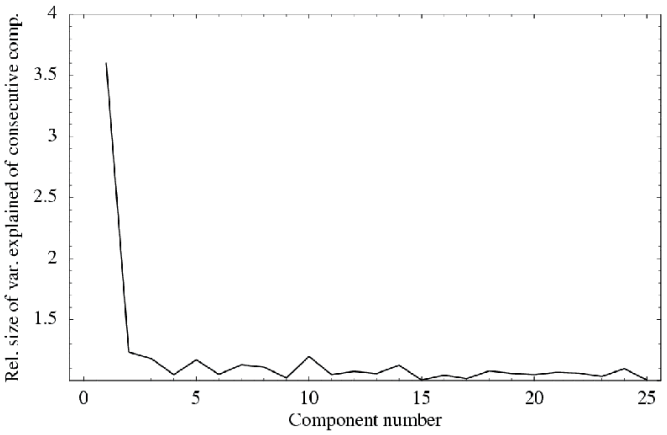


Components of equal the α_c

First principal component

top and bottom categories.	comp.	top and bottom genomes	comp.
cell communication	1.91	Streptomyces coelicolor	0.15
signal transduction	1.91	Streptomyces avermitilis	0.14
regulation of transcription	1.88	Bradyrhizobium japonicum	0.14
electron transport	1.46	Rhizobium loti	0.14
hydrogen transport	0.25	Buchnera aphidicola	-0.19
protein biosynthesis	0.17	Mycoplasma pneumoniae	-0.22
ATP metabolism	0.10	Ureaplasma parvum	-0.24
rRNA modification	0.06	Mycoplasma genitalium	-0.24

Amount of variance captured by the first 25 principal components

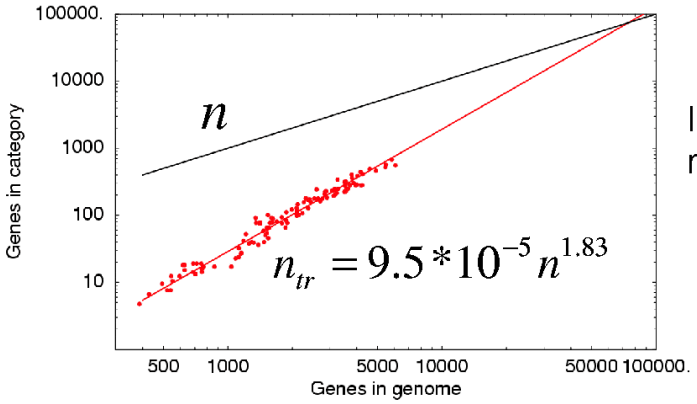


Ratio of the amount of variance captured by components *n* and (*n*+1)

- First principal axis captures 22% of the variance in gene content.
- Second principal axis captures 6% of the variance.
- Third principal axis captures 5% of the variance.
- 55 axis are needed to capture 95% of the variance in the 116 genomes.

Upper bound on genome size

If one naively extends the scaling laws one would eventually have more transcription regulators than there are genes.



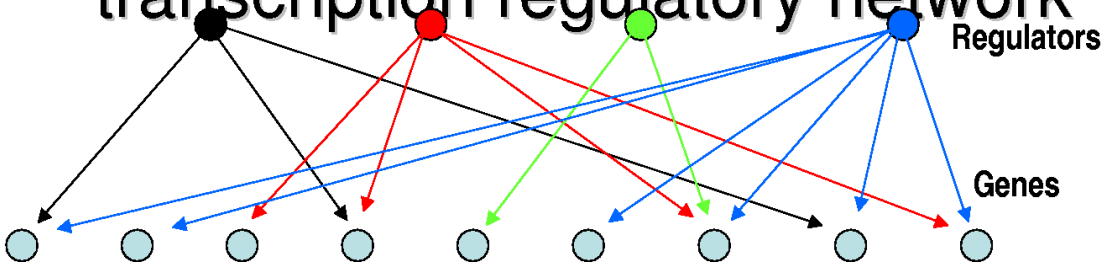
Implies upper bound of n = 70,000

Another bound is obtained by assuming that the number of regulators cannot increase faster than the number of genes

$$dn_{tr} \leq dn \Leftrightarrow A_{tr}(n+1)^{\alpha_{tr}} \leq A_{tr} n^{\alpha_{tr}} + 1$$

This implies an upper bound of n = 34,000. Too large by a factor of 3.

Consequences for the topology of the transcription regulatory network



$\langle r(g) \rangle$ = average number of incoming arrows per gene in genome with *g* genes.

$\langle n(g) \rangle$ = average number of outgoing arrows per regulator in genome with *g* genes.

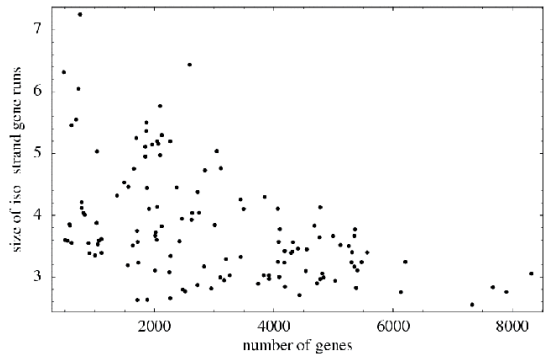
$n_{tr}(g)$ = number of regulators in genome with *g* genes.

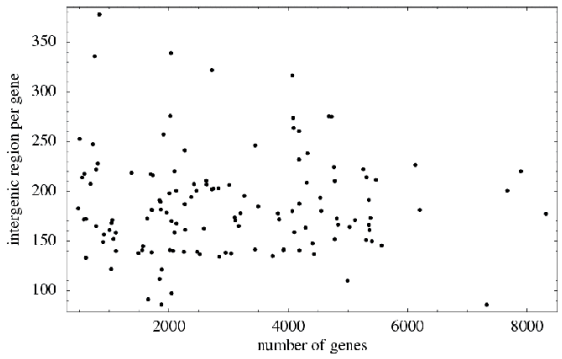
It follows that $$n_{tr}(g)\langle n(g) \rangle = \langle r(g) \rangle g \Leftrightarrow \frac{\langle r(g) \rangle}{\langle n(g) \rangle} \propto g$$

Either genes are regulated by more regulators,
or regulon sizes are dropping as genome size increases

BIOZENTRUM

Operon size and amount of intergenic DNA as a function of genome size.



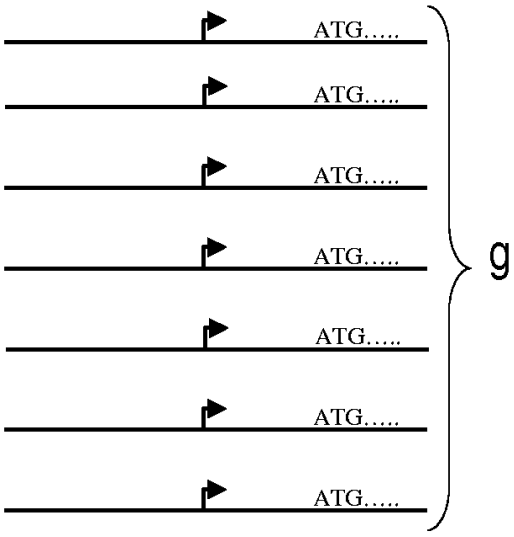


Average length of runs of genes on the same strand as proxy for operon size.
Shows some decrease with genome size (factor of 2-3 over entire genome range).
Amount of intergenic DNA as a function of genome size
Shows no dependence on genome size.

Probably both the number of genes with regulation increases, and average regulon size decreases with genome size.

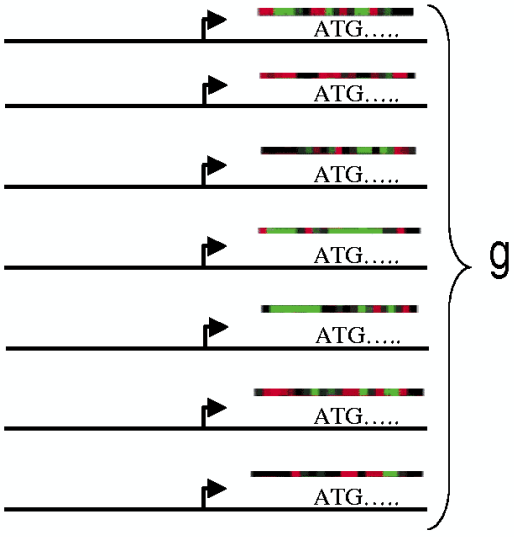
Combinatorial control

- g genes.



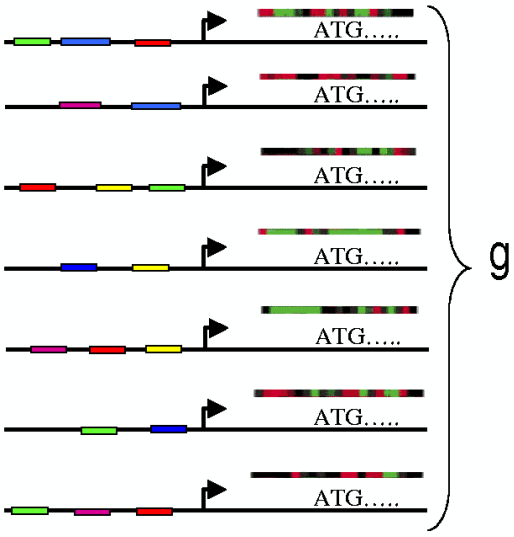
Combinatorial control

- g genes.
- Each with a desired 'expression profile' across different conditions.



Combinatorial control

- g genes.
- Each with a desired 'expression profile' across different conditions.
- To implement these *g* expression profiles one uses binding sites for different combinations of TFs at different genes.



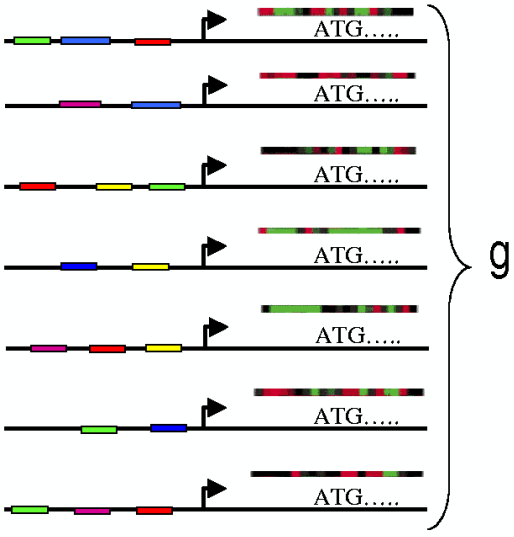
Combinatorial control

- g genes.
- Each with a desired 'expression profile' across different conditions.
- To implement these *g* expression profiles one uses binding sites for different combinations of TFs at different genes.
- If one has *r* different TFs and uses *k* different regulators at each gene one can generate r^k different expression profiles. Thus the number of regulators one needs for *g* different profiles scales as:

$$r \propto g^{1/k}$$

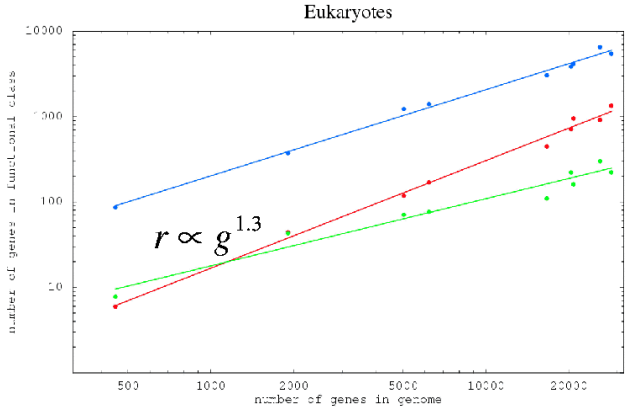
- This is sublinear. We find however that

$$r \propto g^{\alpha} \quad \text{with } \alpha >= 1.3$$



Regulatory 'programs' use genes combinatorially

- Each combination of transcription regulators corresponds to a particular regulatory program.
- If the number of regulatory programs p scales exponentially in the number of regulators r then it must scale *faster* than exponential in the number of genes g.
- Regulators thus combinatorially use genes to implement different regulatory programs.



$$p \propto \exp(r)$$

In Bacteria :

$$p \propto \exp(\lambda g^2)$$

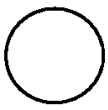
In Eukaryota :

$$p \propto \exp(\lambda r^{1.3})$$

Note: naively this suggest eukaryotes use genes *less* combinatorially

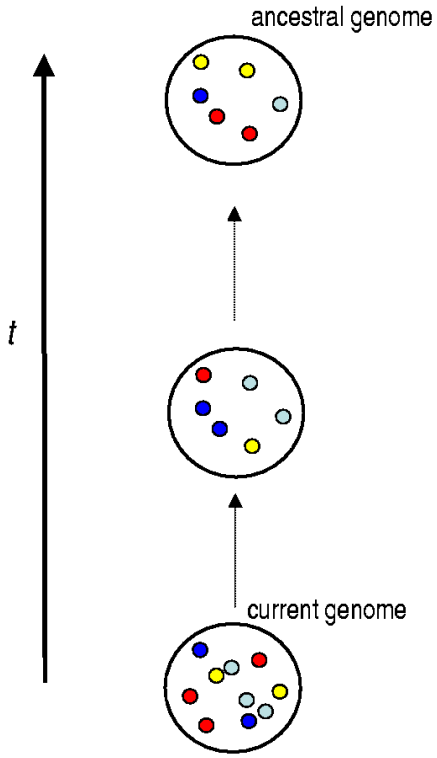
Evolutionary model

- Consider the evolutionary history of a particular genome.
- As one follows the genome back in evolutionary time the total number of genes g and the numbers of genes n_c in different functional categories c have fluctuated in some unknown way.
- We thus have unknown functions of time $g(t)$ and $n_c(t)$ for each category c.
- Assume $g(t)$ and $n_c(t)$ change mainly through gene *duplications* and gene *deletions*.

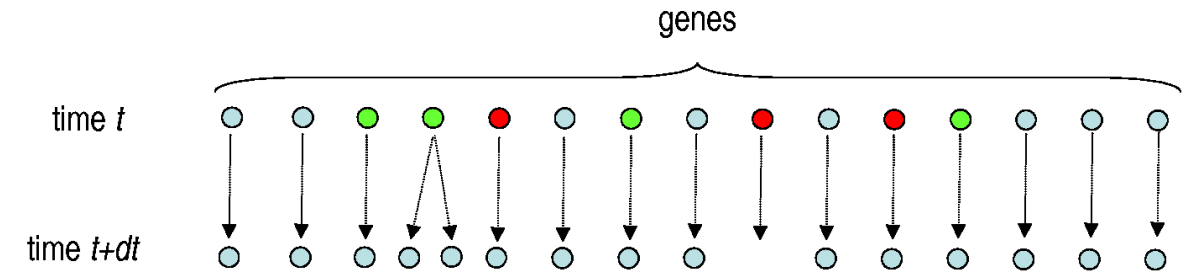

Organism


gene


Genes in different functional categories



Evolutionary model



● **genes that would increase organism fitness when duplicated.** $f_+(t) = 4/15$

● **genes that would increase organism fitness when deleted.** $f_-(t) = 3/15$

- Given a particular time *t* and the environment at that time, there will be a fraction of genes $f_+(t)$ that would benefit the organism when duplicated, and a fraction $f_-(t)$ of genes that would benefit the organism when deleted.
- Assuming the rates λ and δ of *introduction* of duplications and deletions to be the same for all genes, the time dynamics of the total number of genes takes the form:

$$\frac{dg(t)}{dt} = [\lambda f_+(t) - \delta f_-(t)]g(t)$$

Evolutionary model

Evolution equation:

$$\frac{dg(t)}{dt} = [\lambda f_+(t) - \delta f_-(t)]g(t)$$

Formal solution:

$$g(t) = g(0)\exp\left(\int_0^t d\tau[\lambda f_+(\tau) - \delta f_-(\tau)]\right)$$

The integrals are *time averages* over the evolutionary history of the genome:

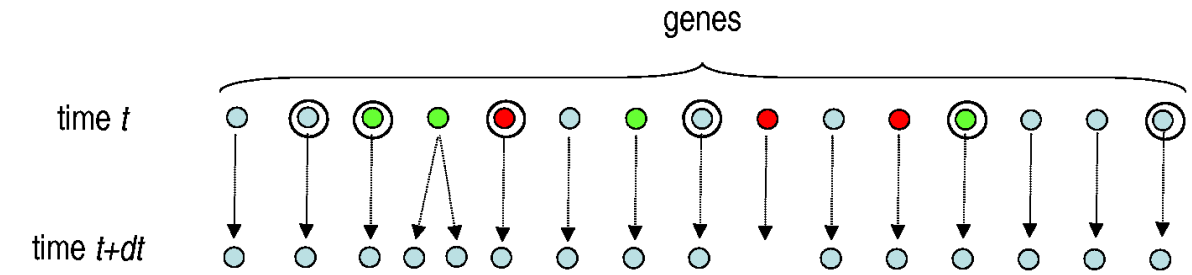
$$\int_0^t d\tau[\lambda f_+(\tau) - \delta f_-(\tau)] = t\lambda\langle f_+\rangle_{gh} - t\delta\langle f_-\rangle_{gh}$$

NOTE: the averages are over the history of a particular genome (i.e. different for each genome).

Solution in terms of genome history averages:

$$g(t) = g(0)\exp\left(\lambda t\langle f_+\rangle_{gh} - t\delta\langle f_-\rangle_{gh}\right)$$

Evolutionary model



- genes that would increase organism fitness when duplicated.
- genes that would increase organism fitness when deleted.
- genes in functional category *c*.

$f_+(t) = 4/15 \quad f_+^c(t) = 2/6$

$f_-(t) = 3/15 \quad f_-^c(t) = 1/6$

- For the genes in category *c* we have an analogous evolution equation in terms of the fractions of genes $f_+^c(t)$ and $f_-^c(t)$ in category *c* that would benefit the organism when duplicated and deleted respectively

$$\frac{dn_c(t)}{dt} = \left[\lambda f_+^c(t) - \delta f_-^c(t)\right] n_c(t)$$

Evolutionary model

Formal solution for genes in category *c*:

$$n_c(t) = n_c(0)\exp\left(t\lambda\left\langle f_+^c\right\rangle_{\text{gh}} - t\delta\left\langle f_-^c\right\rangle_{\text{gh}}\right)$$

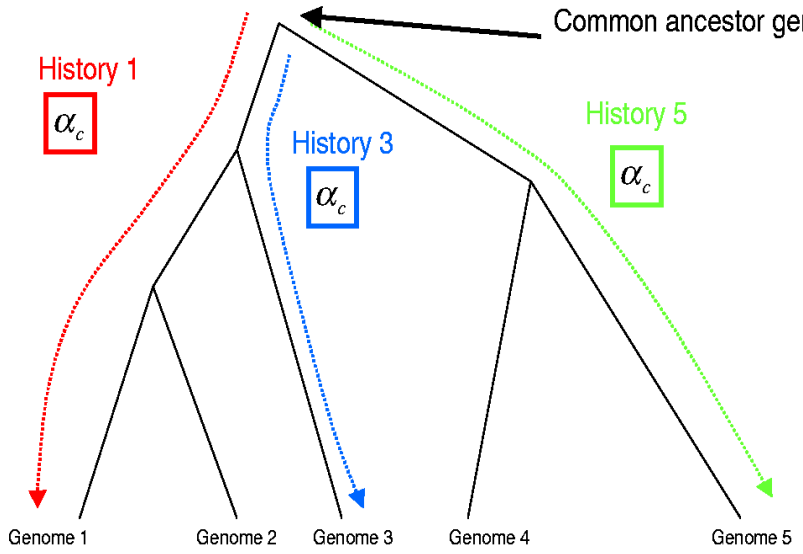
Using the solution for the total number of genes

$$g(t) = g(0)\exp\left(t\lambda\left\langle f_+\right\rangle_{\text{gh}} - t\delta\left\langle f_-\right\rangle_{\text{gh}}\right)$$

We can expression $n_c(t)$ in terms of $g(t)$

$$n_c(t) = n_c(0)\left[\frac{g(t)}{g(0)}\right]^{\alpha_c} \qquad \alpha_c = \frac{\lambda\left\langle f_+^c\right\rangle_{\text{gh}} - \delta\left\langle f_-^c\right\rangle_{\text{gh}}}{\lambda\left\langle f_+\right\rangle_{\text{gh}} - \delta\left\langle f_-\right\rangle_{\text{gh}}}$$

Evolutionary histories of multiple genomes



For each genome separately holds:

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

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

α_c = α_c = α_c

The exponents correspond to evolutionary constants.

Evolutionary Susceptibilities

α_c = α_c = α_c

$$\alpha_c = \frac{\lambda \left\langle f_+^c \right\rangle_{\text{gh}} - \delta \left\langle f_-^c \right\rangle_{\text{gh}}}{\lambda \left\langle f_+ \right\rangle_{\text{gh}} - \delta \left\langle f_- \right\rangle_{\text{gh}}}$$

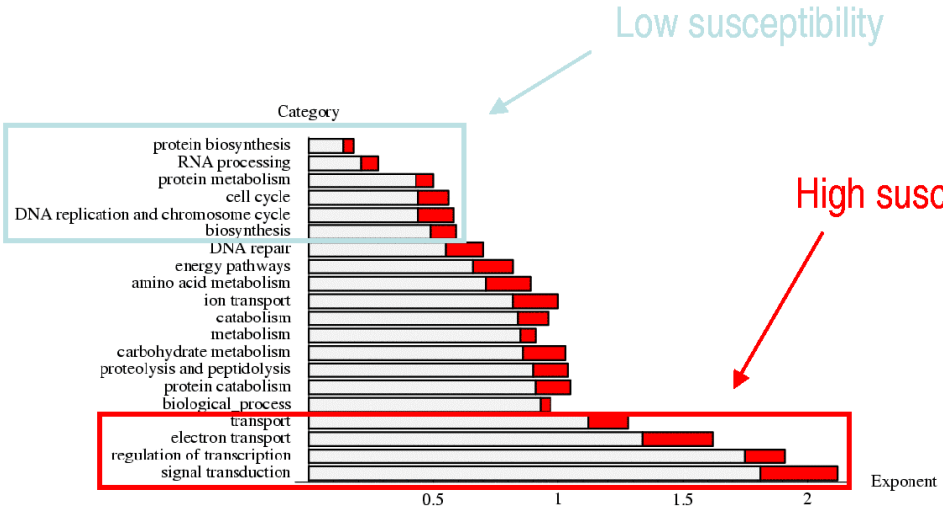
is the same for every genome history.

Sufficient condition $\left\langle f_+^c \right\rangle = \alpha_c \left\langle f_+ \right\rangle, \quad \left\langle f_-^c \right\rangle = \alpha_c \left\langle f_- \right\rangle$

Genes in categories with high exponent α_c are more evolutionary *susceptible*:
as the environment changes it is more often beneficial to duplicate or delete these genes.

Evolutionary susceptibilities

$$\left\langle f_+^c \right\rangle = \alpha_c \left\langle f_+ \right\rangle, \quad \left\langle f_-^c \right\rangle = \alpha_c \left\langle f_- \right\rangle$$

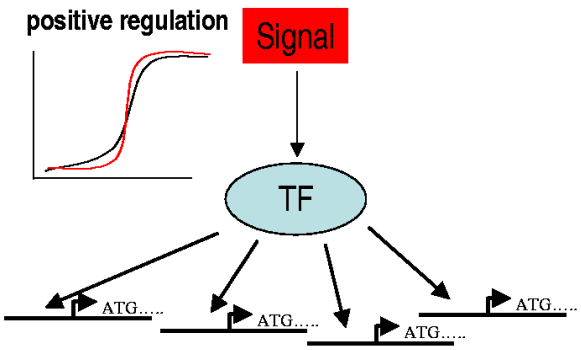


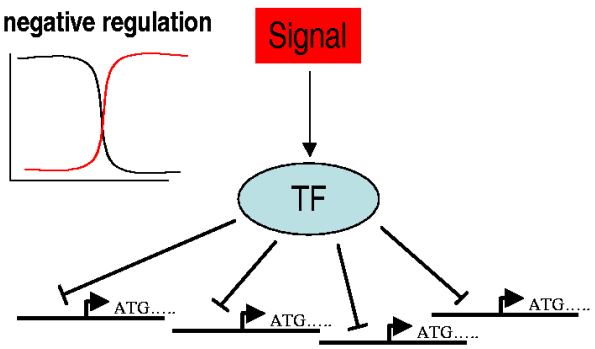
Speculations on the quadratic scaling of transcription regulators

For transcription regulators:

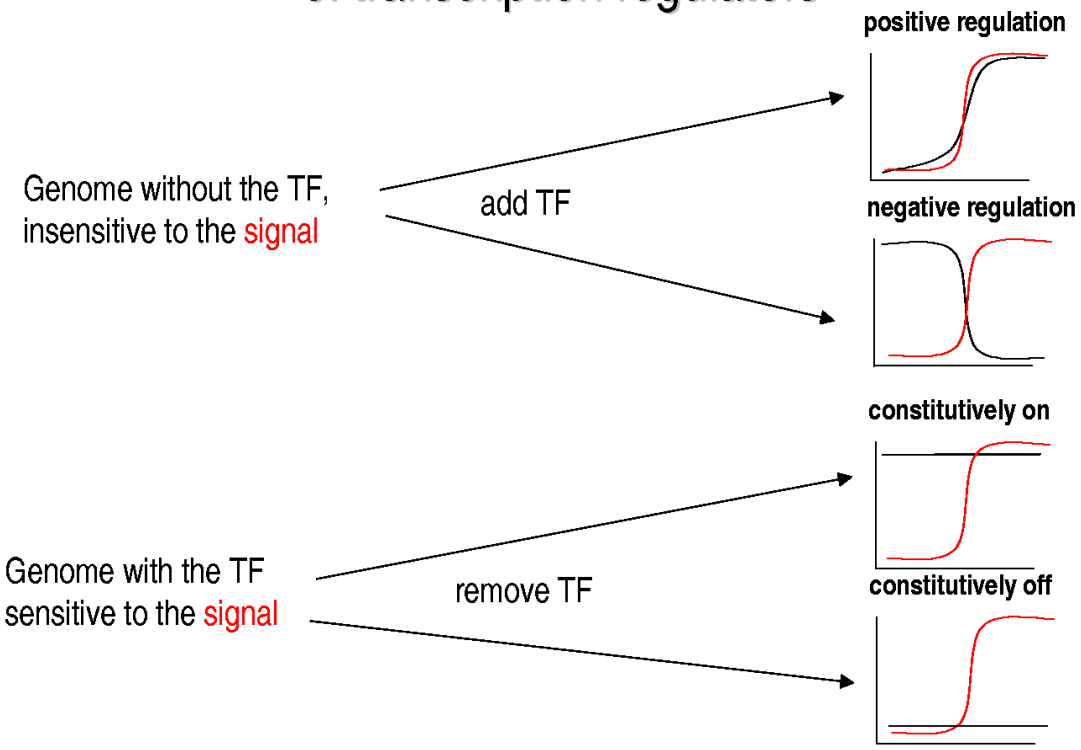
$$\left\langle f_c^+ \right\rangle \approx 2\left\langle f^+ \right\rangle, \qquad \left\langle f_c^- \right\rangle \approx 2\left\langle f^- \right\rangle$$

Proposal: factor 2 stems from switch-like function of regulators.
A transcription factor can respond to the same signal and regulate the same set of genes in essentially two ways:





Speculations on the quadratic scaling of transcription regulators



Speculations on the quadratic scaling of transcription regulators

- Because of their switch-like function, TFs can be removed or added to the regulatory network in 2 distinct ways.
- This is not true for metabolic genes and structural genes.
- Therefore:

$$\left\langle f_c^+ \right\rangle \approx 2\left\langle f^+ \right\rangle, \qquad \left\langle f_c^- \right\rangle \approx 2\left\langle f^- \right\rangle$$

- The same argument applies to signal-transduction genes.
- Indeed these also scale quadratically with number of genes in bacterial genomes.

- Nontrivial exponents for other functional categories cannot be explained by such arguments.