genome-wide studies using biocyc

- 0 | background
- 1 | functional profiling
- 2 | partial reconstruction
- 3 | functional associations
- 4 | evolutionary profiling
- 5 | network inference
- 6 | ancestral reconstruction

christos a. ouzounis

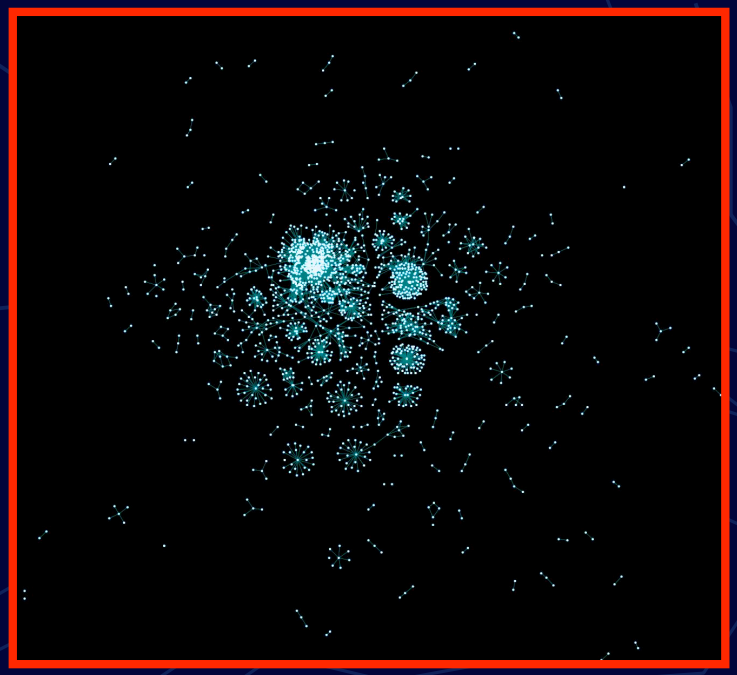
email: cao@ebi.ac.uk

computational genomics group @ ebi

url: cgg.ebi.ac.uk

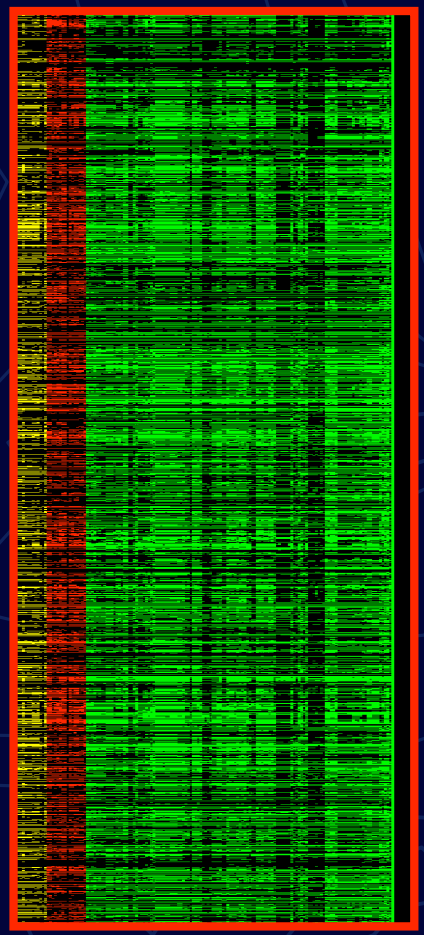
networks: function

- Metabolic pathways
  - Profiling metabolic maps
    - Genome Res 10:568, 11:1503
  - Conservation of metabolism
    - Genome Res 13:422
  - Automatic pathway reconstruction, MjCyc
    - Archaea 1:223
  - Partial-genome reconstruction
    - J Bioinfo Comput Biol 2:589
  - Pathways for 160 genomes
    - Nucl Acids Res 33:6083
- Functional modules, interaction networks
  - Detection of functional modules by clustering
    - Proteins 54:49
  - Ancestral state reconstruction of interactions
    - Mol Biol Evol 21:1171
  - Exponential distribution of interactions
    - Mol Biol Evol 22:421
  - Functional associations in gene networks
    - Alonso Cases & CAO, unpublished:2004



genomes: evolution

- Sequence clustering
  - All genomes
    - Bioinformatics 19:1451
  - All phylogenetic profiles
  - All protein families
    - Nucl Acids Res 31:4632
  - All sequence similarities, ortholog clusters
  - All genome based trees
    - Nucl Acids Res 33:616
  - All gene fusions
    - Genome Biol 2:r0034.1
- Genome evolution patterns
  - Comparison of different cellular processes
    - Trends Microbiol 11:248
  - Reconstruction of funtional modules from genome constraints
    - Proc Natl Acad Sci USA 100:15428
  - Ancestral state reconstructions with loss and HGT
    - Genome Res 13:1589
  - Functional content of last universal common ancestor
    - Res Microbiol in press:2005

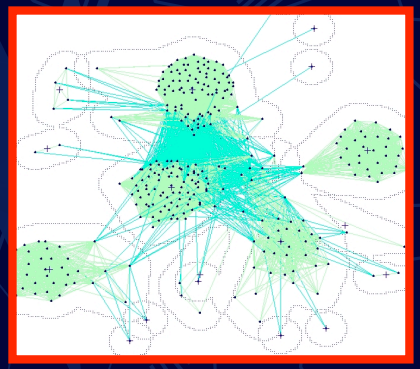


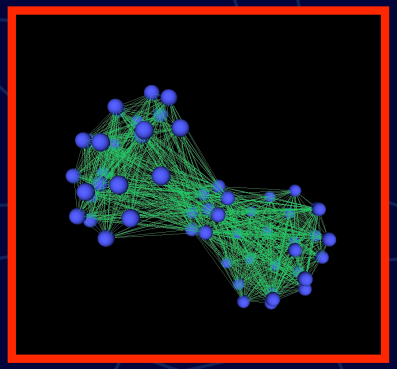
Enright • Kunin • Kreil • Cases • Darzentas • Iliopoulos • Audit • Goldovsky

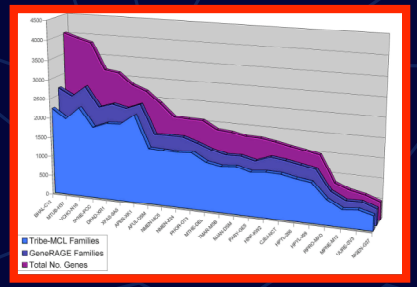
algorithms

- CAST for low-complexity masking
  - Bioinformatics 16:915
- geneRAGE for domain clustering
  - Bioinformatics 16:451
- TRIBE-MCL for rapid clustering
  - Nucl Acids Res 30:1575
- bioLayout for similarity visualization
  - Bioinformatics 17:853
- geneTRACE for ancestral reconstruction
  - Bioinformatics 19:1412
- TextQuest for document clustering
  - Pac Symp BioComp 6:384

```
maine> cast SRm160.f
>SRm160
MDAGFFRGTSAEQDNRFSNKQKKLLKQLKFAECLEKKVDMSKVNLEVIKPWITKRVTEIL
GFEDDVVIEFIFNQLEVKNPDSKMMQINLTGFLNGKNAREFMGELWPLLLSAQENIAGIP
SAFLXLXXXXIXQRQIXQXXLASMXXQDXDXDXRDXXXXXXXXXXXXXXXXXPXXXXXXP
XPXXXXXPVXXEXXXXHXXPXHXTXXXXPXPAPEXXEXTPELPEPXVXVXEPXVQEATX
TXDILXVPXPEPIPEPXEPXPEXNXXXEQEXEXTXPXXXXXXXXXXXTXXXXPXHTXPXX
XHXXDXMYXXXXXXXXXXXXXXXXXXXTXXXXMXXXXXHXXXXXPVXXXXXXXXAXLXGXXXX
XXXXXXXXXXXXXXXXXTXXXXXXTXXLXXXAXXXXXXHXXXXXATXXXTXDXXTXQQXNX
TXXXXVXVXPGXTXGXVTXHXGTEXXEXPXPAPXPXXVELXEXEEDXGGXMAAADXVQQX
XQYXXQNQQXXXDXGXXXXXEDEXPXXXHVXNGEVGXXXXHXPXXXAXPXPXXXQXETXP
...
```



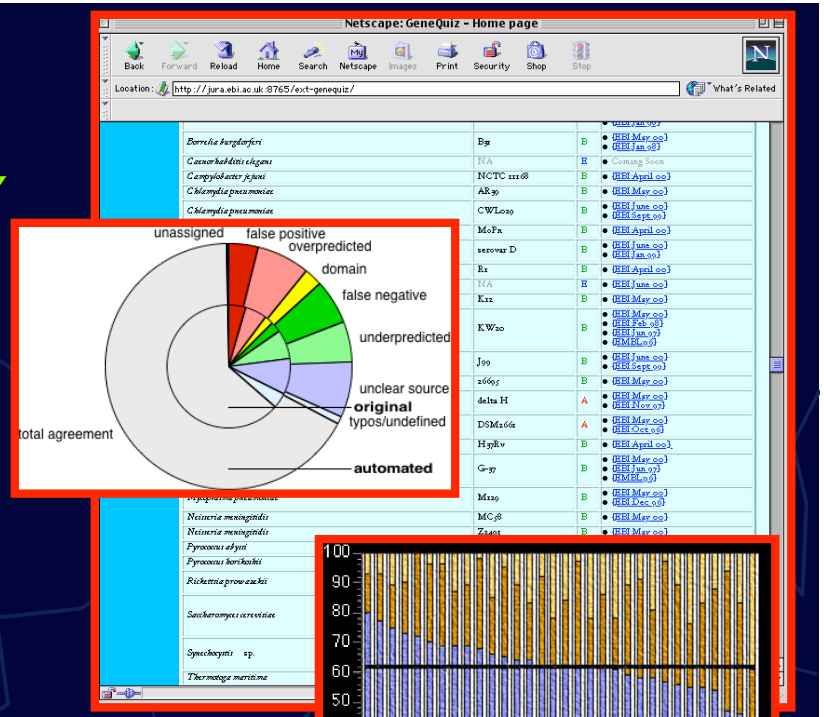


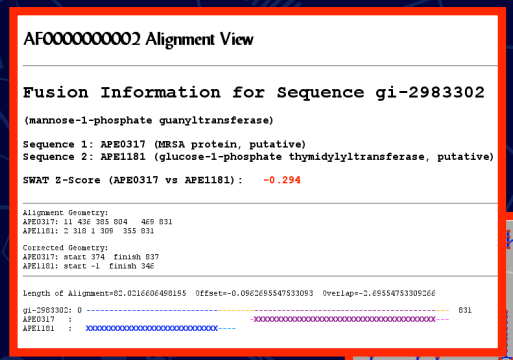


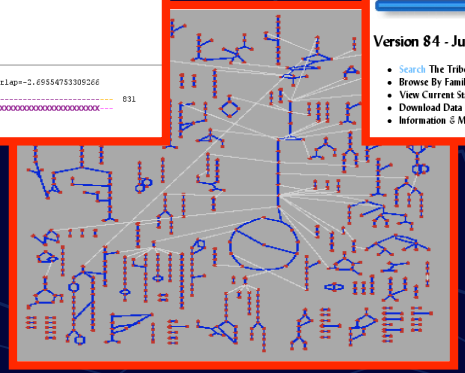
Enright • Kunin • Tsoka • Audit • Janssen • Iliopoulos • Darzentas • Goldovsky

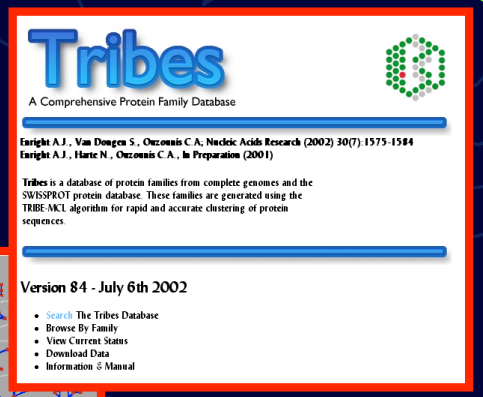
databases

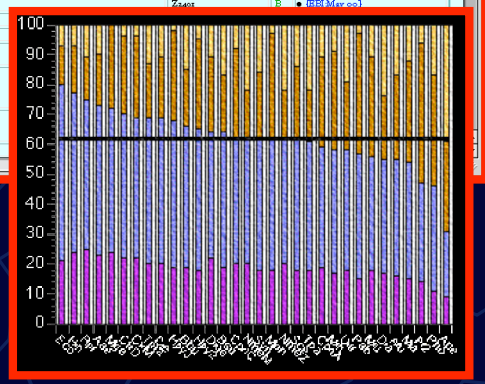
- **geneQUIZ** for annotation
  - Bioinformatics 15:391
- **CoGenT** for tracking genomes
  - Bioinformatics 19:1451
- **CoGenT++** for genome analysis
  - Bioinformatics 21:3806
- **Tribes** for protein families
  - Nucl Acids Res 31:4632
- **AllFuse** for gene fusions
  - Genome Biol 2:r0034.1
- **BioCyc** for metabolic pathways
  - Nucl Acids Res 33:6083

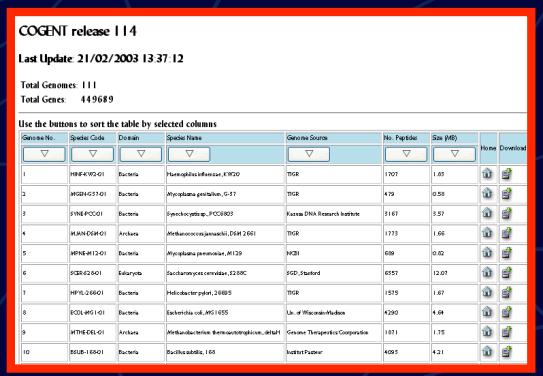






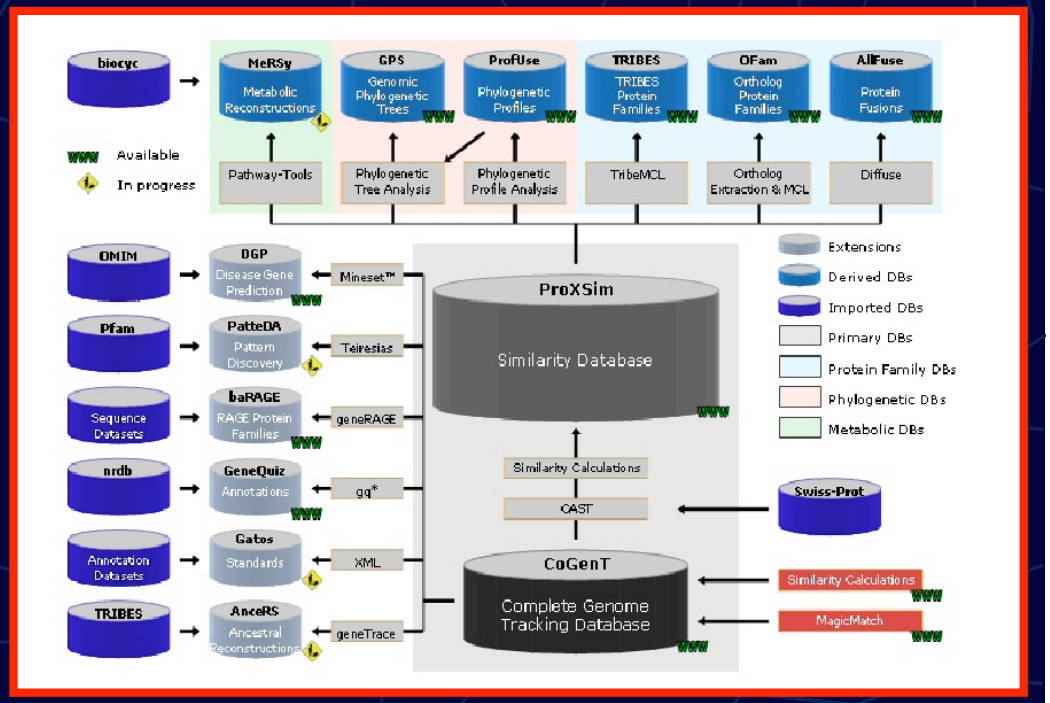






genome analysis in COGENT

- Closely following timed releases
- COGENT/++ as resource
  - Bioinformatics 21:3806
  - COGENT: complete genomes
  - ProXSim: similarity information
  - AllFuse: interactions by fusion
  - Ofam: putative orthologs
  - TRIBES: putative homologs
  - ProfUse: phylogenetic profiles
  - GPS: genome phylogenetic trees
  - MeRSy: predictions from BioCyc
  - GeneQuiz: automatic annotation
- Queriable by:
  - sequence, via BLAST
  - identifier, via MagicMatch
- MagicMatch
  - >5 million links to major databases
    - Bioinformatics 21:3429
- 12X UniProt in public-access data



a sense of scale

Genomes	200 (243)
Genes	751,742 (915,554)
Annotations	359,482
Similarities	384,579,409
Phylogenetic profiles	181,986
Families	82,692
Interactions	2,192,019
Pathways	> 25,000

genome analysis with biocyc

- Examples of studies using the BioCyc environment
  - Brief descriptions of what has been done
  - Discussion of what else can be done…
- Three cases of single-species analysis
- Three cases of multiple-species analysis
- All these computations external to COGENT
  - … except last (most complex) case, where the power of combining both systems is showcased

1 | functional profiling

- Analysis of known metabolic complement of *E. coli*
  - correlations between compounds, reactions, enzymes, pathways
    - Genome Res 10:568
- Identified promiscuous reactions in terms of enzymes and pathways (example in figure below)
- Model case study for comparative genomics, still unexplored territory
- Today possible, for curated or automatically generated maps

Global Properties of the Metabolic Map of *Escherichia coli*

Christos A. Ouzounis[1,3] and Peter D. Karp[2]

[1]Computational Genomics Group, The European Bioinformatics Institute, EMBL Cambridge Outstation, Cambridge, CB10 1SD, UK; [2]Bioinformatics Group, AI Center, SRI International, EK223, Menlo Park, California 94025 USA

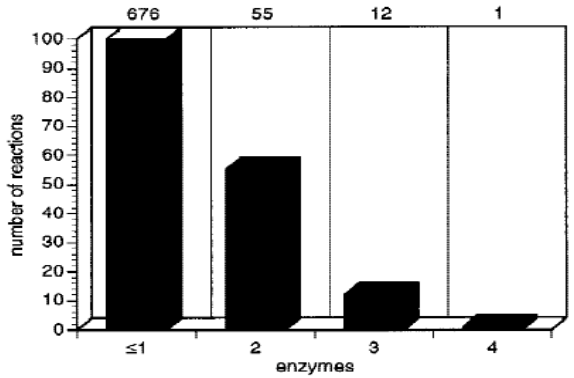


Figure 6 Diagram showing the number of reactions that are catalyzed by one or more enzymes. Most reactions are catalyzed by one enzyme, some by two, and very few by more than two enzymes.

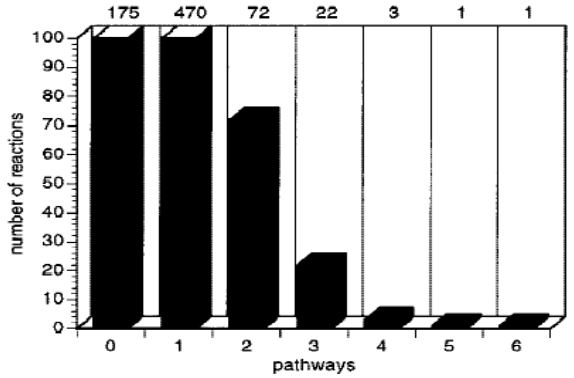
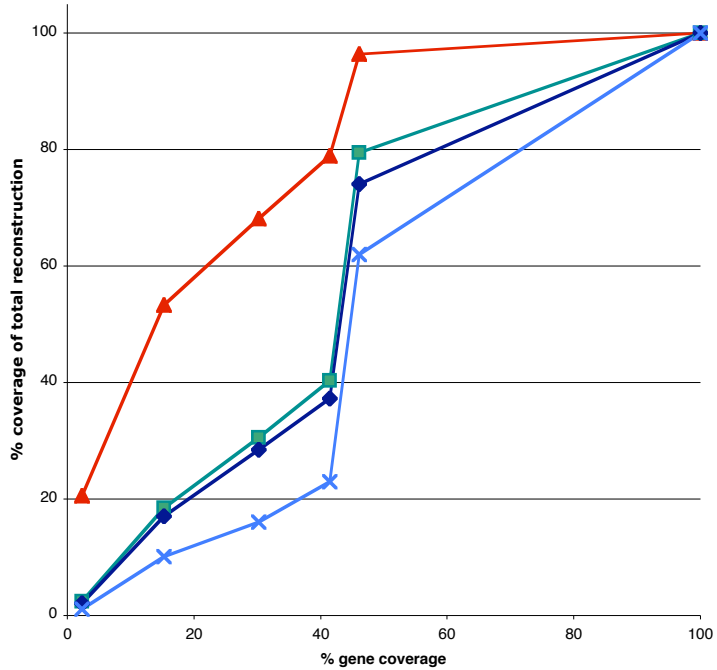


Figure 8 Diagram showing the number of reactions that participate in one or more pathways.

sequenced to date have identified virtually no multi-

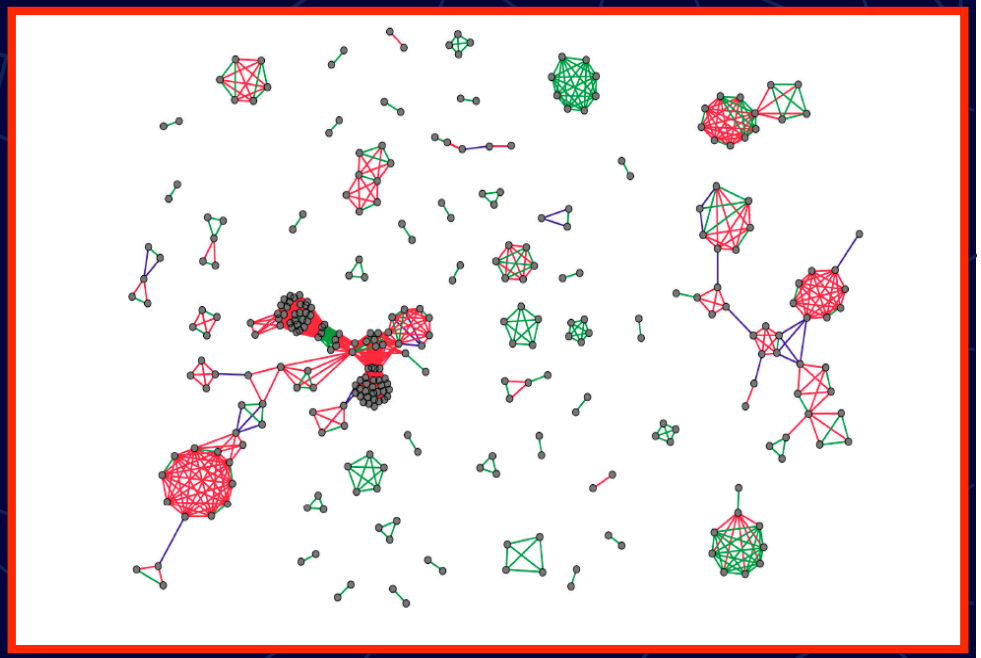
2 | partial reconstruction

- Benchmark on partial genomes
  - Simulated data from the *S. pombe* project
- Robustness and predictive power of reconstruction
  - J Bioinfo Comput Biol 2:589
  - Build total map from complete genome as reference
  - Take datasets from time points during genome project
  - Count unique reactions, enzymes and pathways and compare with reference
- Detection with partial genomes
  - Phase transition at 50% ?
  - Pathways detected at various degrees, without closure
  - Alternatively, build all maps
  - Initial idea, requires more comparative work



3 | functional associations

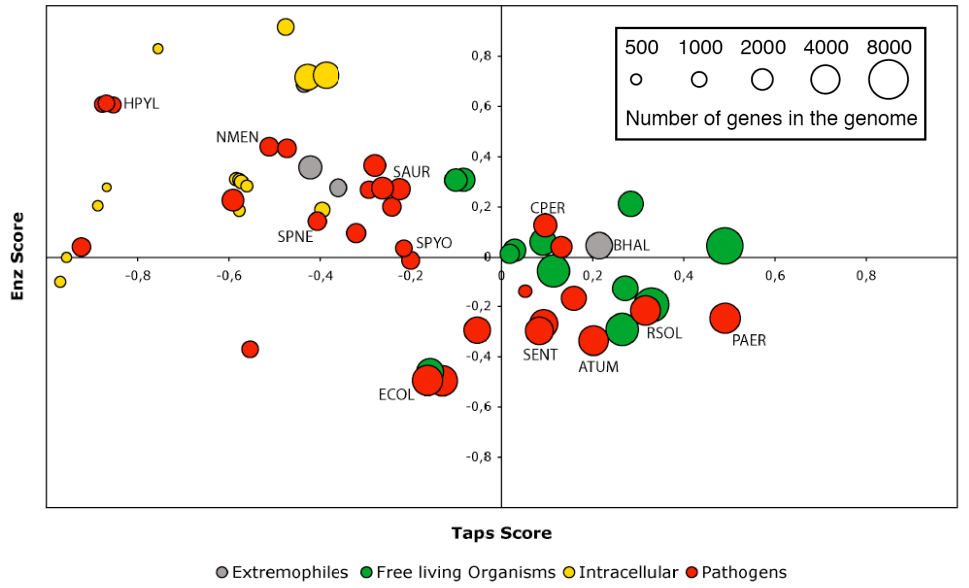
- What are the functional relationships in real gene networks?
  - unpublished:2004
- Analysis of *E. coli* networks using OperonDB and BioCyc
  - detected all networks and overlayed functional information
- Example: what is the relationship between pathways and transcription units in gene network?



  - red: same pathway
  - blue: same transcription unit
  - green: both same pathway and transcription unit
  - if same TU, then 95% in same pathway
- Applications in functional genomics, e.g. microarray-based networks

4 | evolutionary profiling

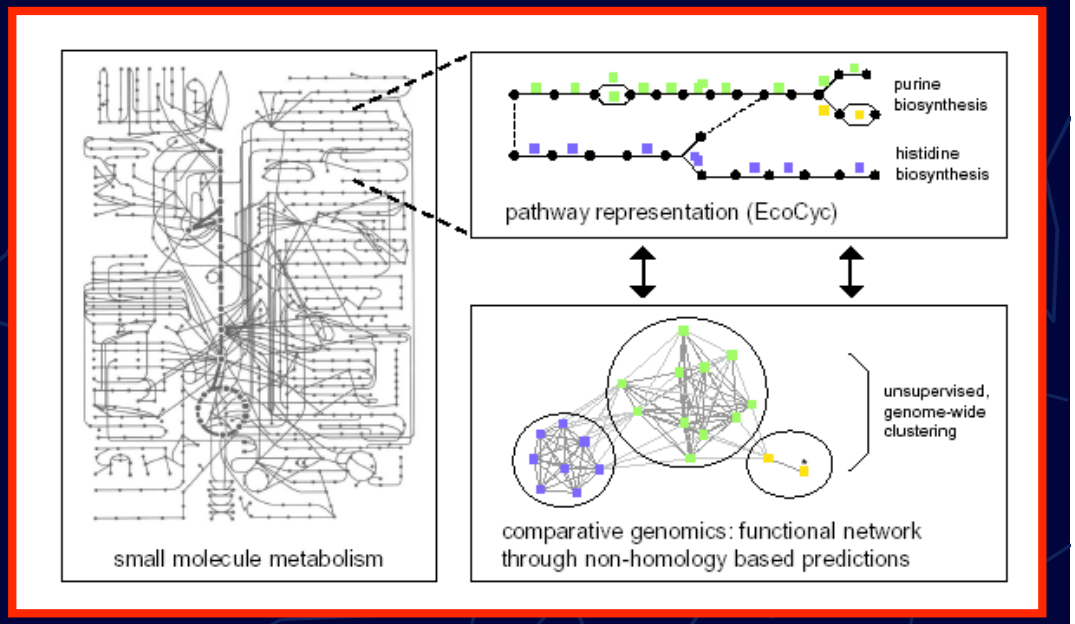
- Correlate genome size and environmental niche with enzymes and transcription-associated proteins (4D)



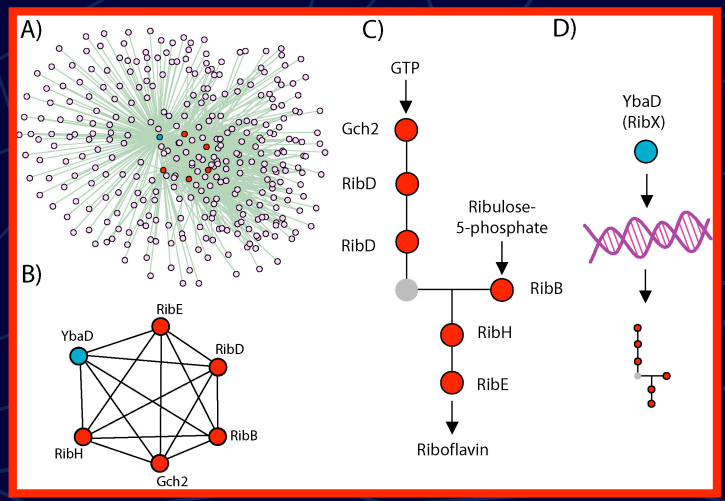
- Larger genomes have more TAPs
- Free-living organisms have more TAPs, intracellular organisms have less
- No clear patterns for enzymes
- Small genomes somewhat richer in enzymes
- Indirect evidence that metabolism is more conserved
  - Trends Microbiol 11:248

5 | network inference

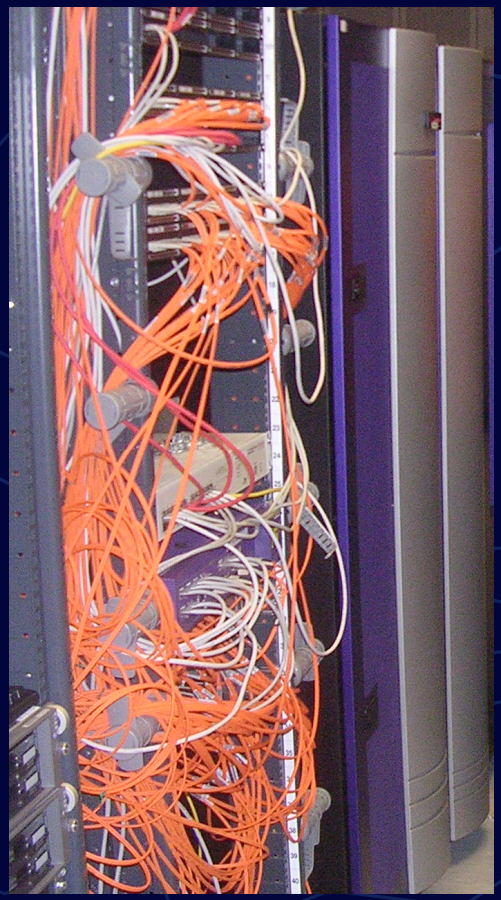
- Integrate all context-based predicted protein interactions
- Cluster protein network (various methods)
- Compare predicted cluster vs. known pathway assignments



- 74% of 583 metabolic enzymes cluster in 119 modules with 84% average pathway specificity and 49% average sensitivity
  - Proc Natl Acad Sci USA 100:15428
- Much of bacterial metabolism is encoded in genome structure!
- Novel functions predicted



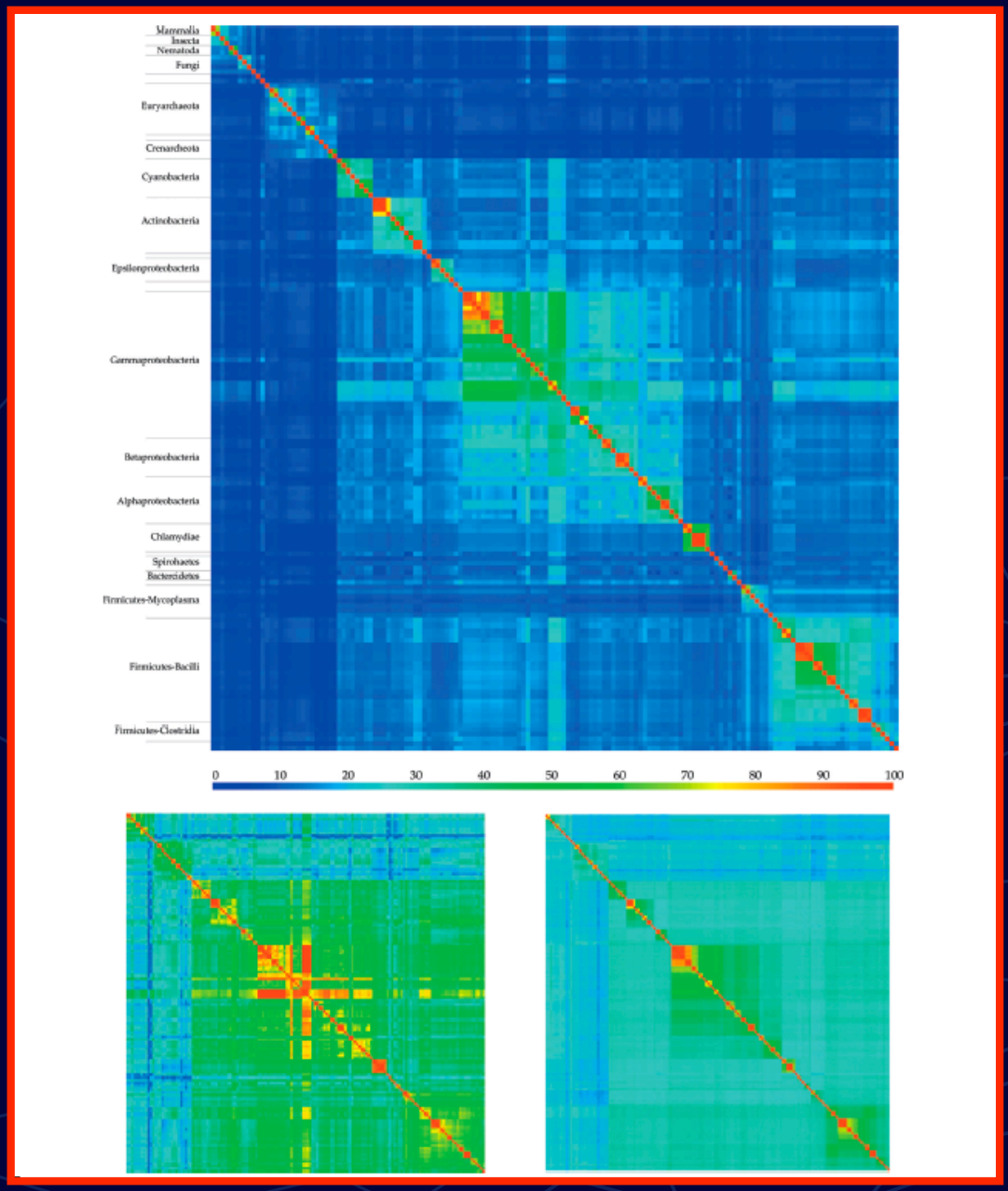
more complicated stuff...



Kunin • Ahrén • Goldovsky • Janssen

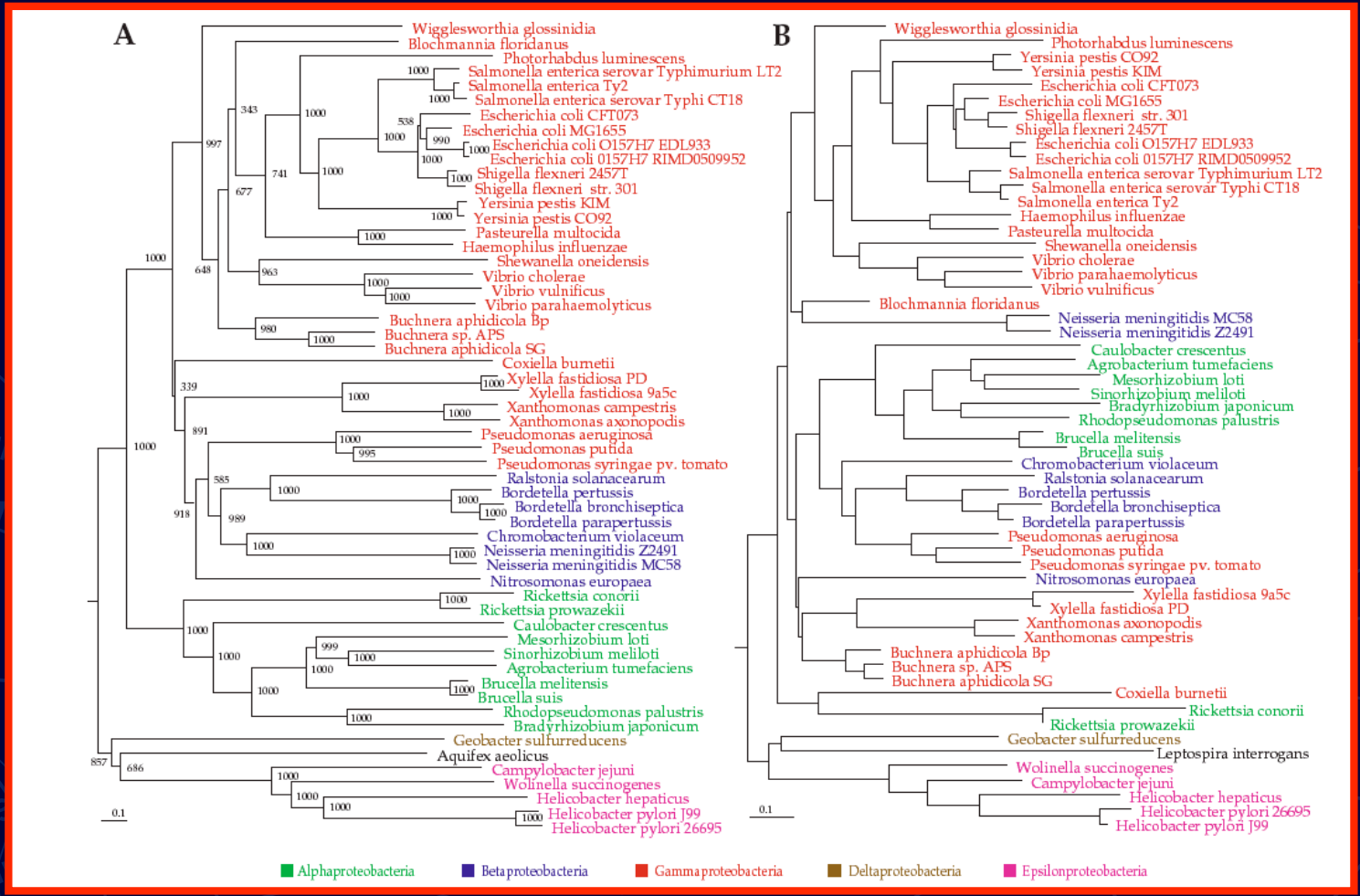
genome similarity maps

- Genome trees from both gene content and average sequence similarity
  - 153 genomes (in publication), >200 now
  - >25 M pairs
    - Nucl Acids Res 33:616



- $\sum (A,B) \neq \sum(B,A)$
- $S(A,B) = \min(\sum(A,B),\sum(B,A))$
- D=D1 or D2
- $D1=1-S/\min((\sum(A,A), \sum(B,B))$
- $D2=-\ln(S/(\sqrt{2} * \sum(A,A) * \sum(B,B) / \sqrt{(\sum(A,A)^2 + \sum (B,B)^2)}$

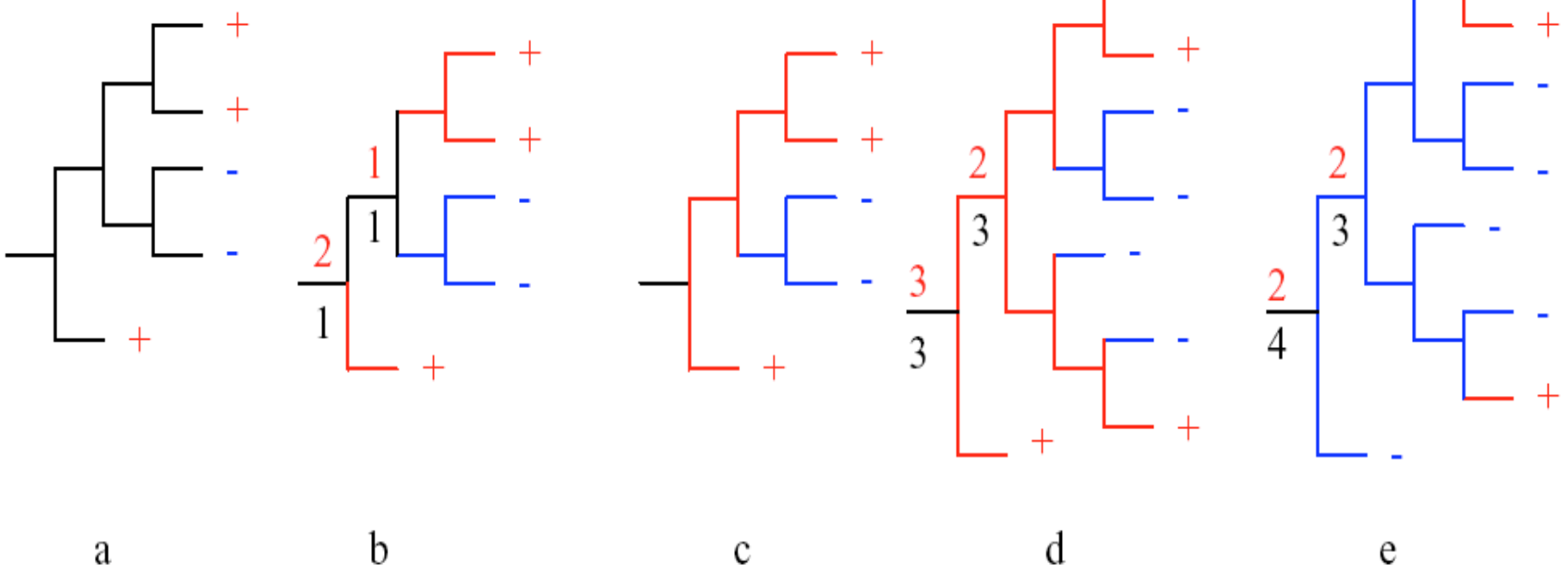
genome conservation trees



ancestral gene content

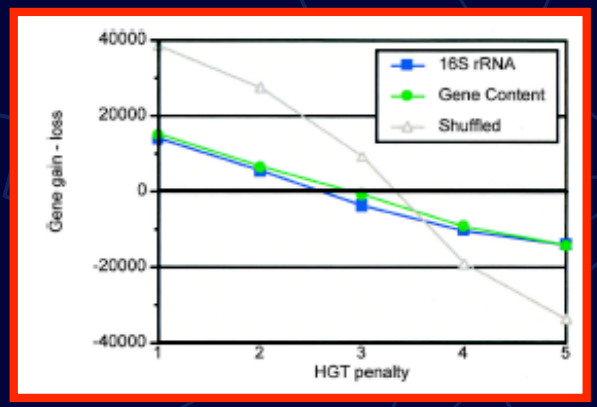
- Family distributions on a guide tree
  - Can be rRNA tree, gene content, genome conservation
  - Original implementation on rRNA trees
  - Terminal nodes (extant genomes) contain gene/protein families
- Assumptions
  - When most clade members contain representative, indication of vertical descent
  - When few clade members do not contain representative, widely distributed in neighborhood, indication of gene loss
  - Interspersed family distribution across remote clades indicative of horizontal gene transfer
- The GeneTrace algorithm
  - Bioinformatics 19:1412
  - Input: Phylogenetic profiles of families and guide tree
  - Inner nodes represent ancestral organisms (states)
  - Start at terminal nodes towards the root
  - Scoring of gain/loss of parental nodes equal to sum of daughter nodes
  - Scores transformed to assignments of presence/absence of EACH family
  - Tags certain families as candidates for horizontal gene transfer

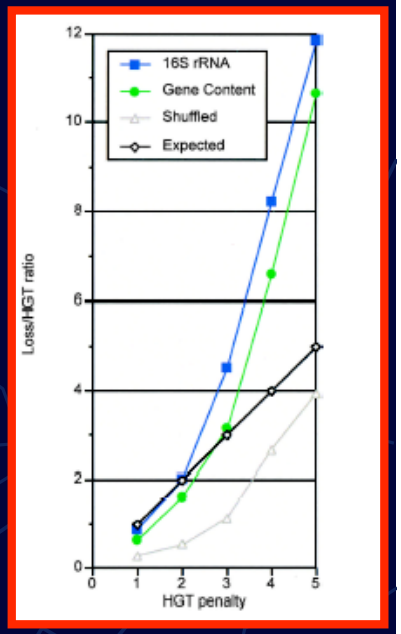
ancestral gene content

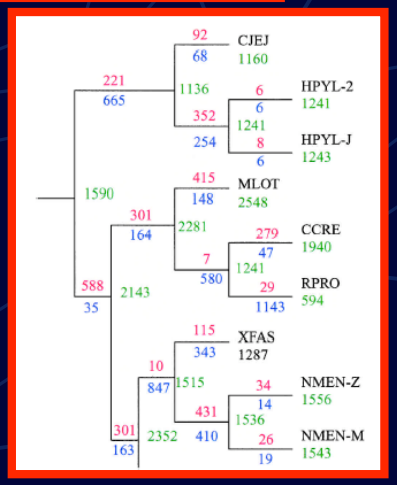


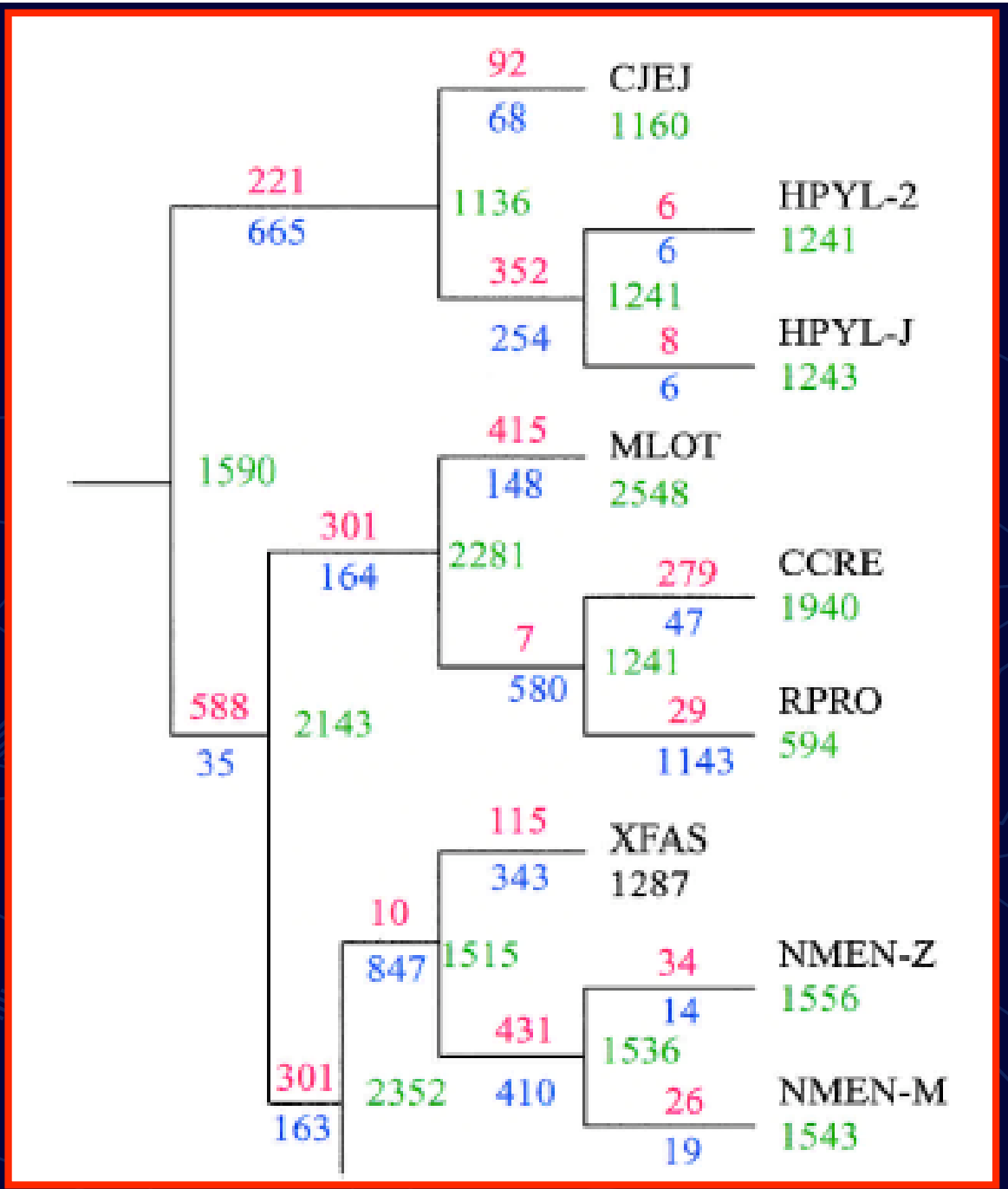
evolutionary scenarios

- Overlay family profiles on trees
  - Play parsimonious guesswork game of gain (genesis + HGT) and loss
- Parameter calibration
  - HGT's 'expected relative frequency' ≈ loss/HGT
  - On average: gain ≈ loss
    - Genome Res 13:1589
- Relative contributions
  - loss:genesis:HGT = 3:2:1
- Reconstruct ancestral states
  - Families
  - Interactions
    - Mol Biol Evol 21:1171









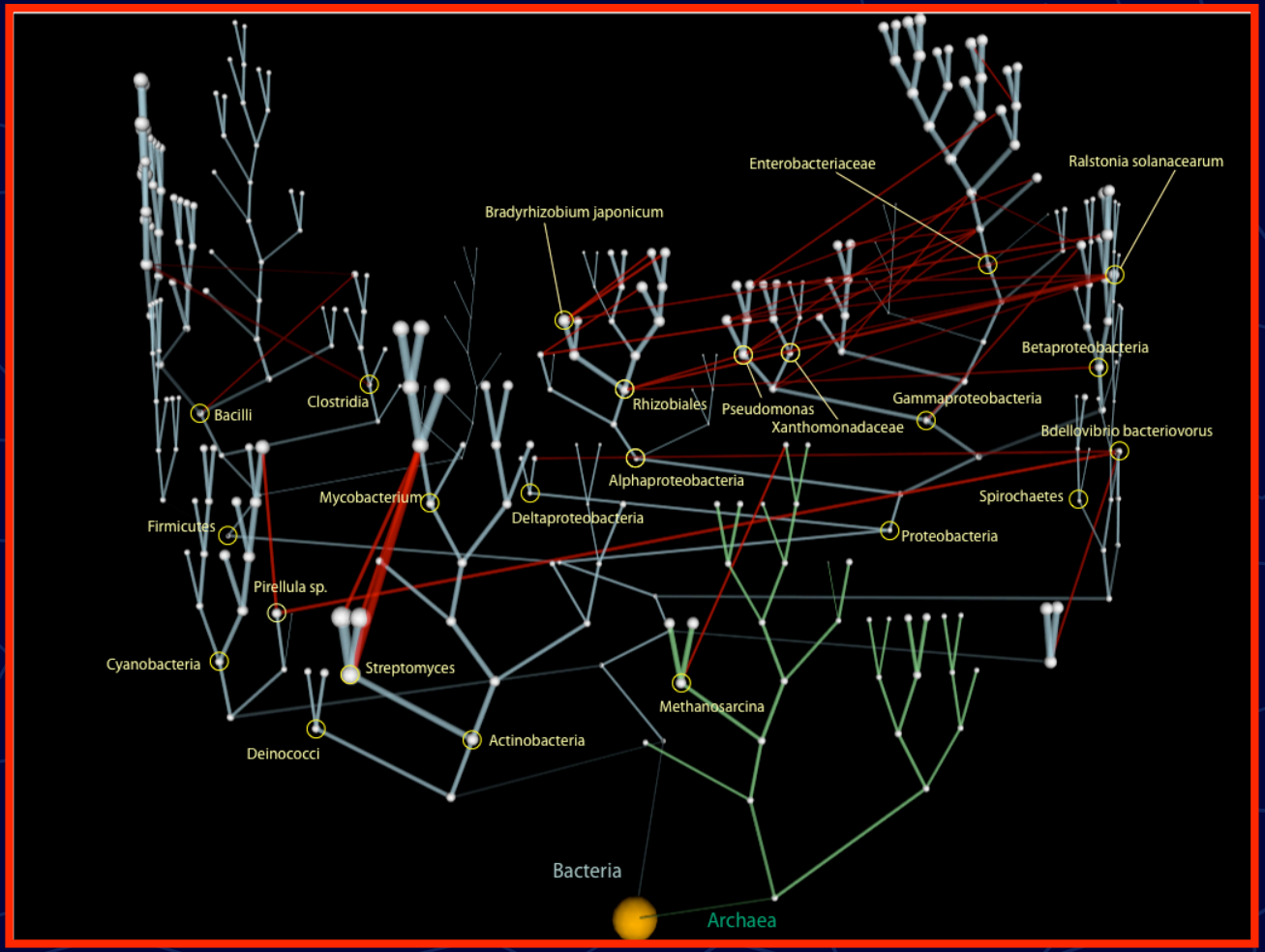
tree or network?

- Tree representations of species relationships
  - Drawback: dealing only with vertical inheritance
  - Data:
    - 184 genomes
    - Bidirectional best hits (putative orthologs…)
    - 3 different guide trees: gene content, average similarity, genome conservation
  - HGT champions

Organism	Average ortholog similarity	Gene content	Genome conservation	STRING
Pirellula sp.	2	1	1	Absent
Bradyrhizobium japonicum	3	3	2	4
Erwinia carotovora	5	2	4	Absent
Clostridium acetobutylicum	4	4	10	5
Chromobacterium violaceum	6	10	9	Absent

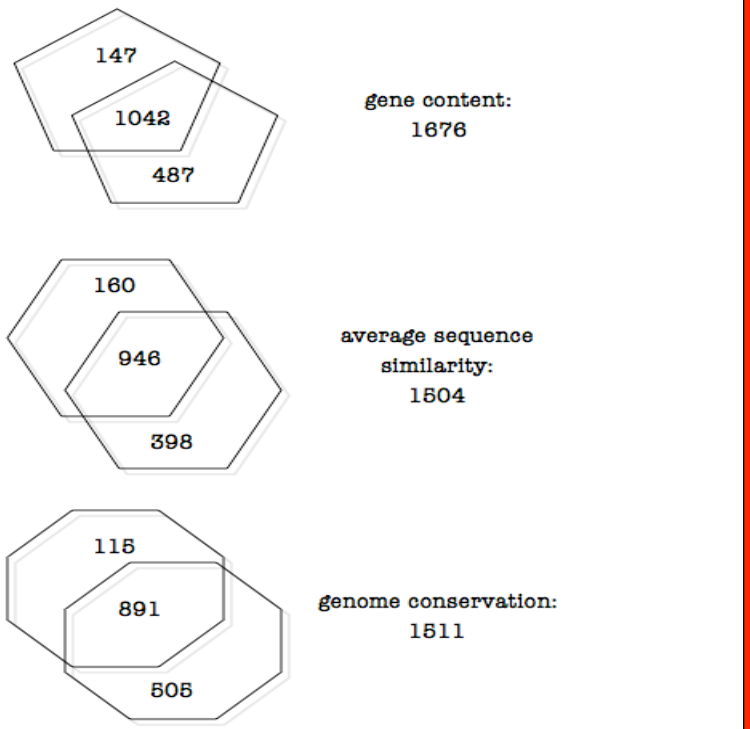
network of life

- Gene content reconstruction: ancestral states
- Both current and ancient genomes
  - HGT events on tree
  - Not directional!??
- Power-law patterns:
  - Number of HGT's between any nodes
  - HGT network connectivity
    - Genome Res 15:954

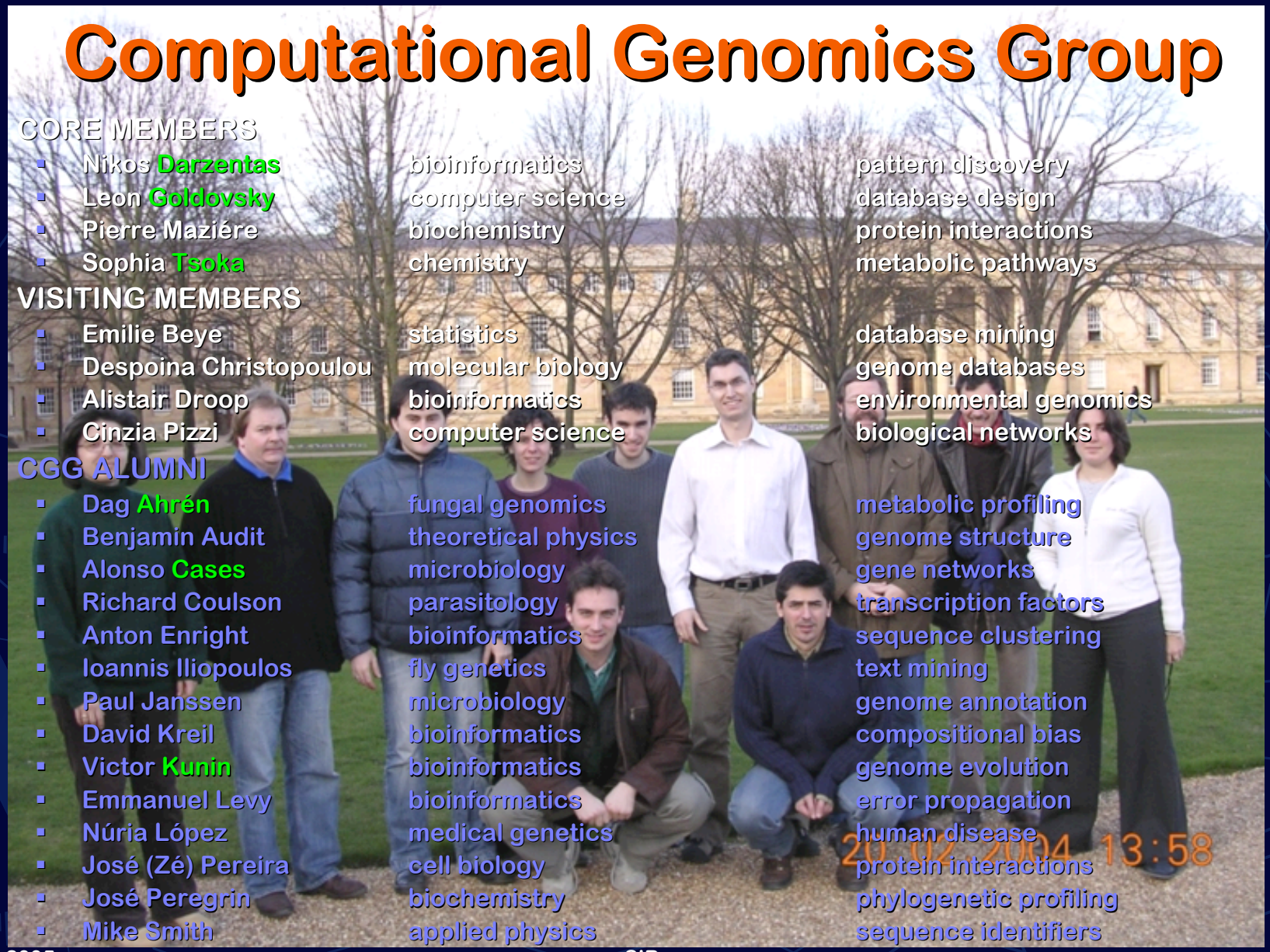


6 | ancestral reconstruction

- The gene content of the Last Universal Common Ancestor
  - Reconstruction of ancestral states for 37,402 families from 184 genomes
- Depending on method: 1,006 to 1,189 protein families
  - Extremely robust estimates, based on ancestral state reconstruction, parsimony
- Notion of similarity to extant genomes of obligate parasites not supported (also known as the 'minimal genome hypothesis')
  - Res Microbiol in press:2005 (special issue on Exobiology)



Computational Genomics Group



CORE MEMBERS		
Nikos Darzentas	bioinformatics	pattern discovery
Leon Goldovsky	computer science	database design
Pierre Maziére	biochemistry	protein interactions
Sophia Tsoka	chemistry	metabolic pathways
VISITING MEMBERS		
Emilie Beye	statistics	database mining
Despoina Christopoulou	molecular biology	genome databases
Alistair Droop	bioinformatics	environmental genomics
Cinzia Pizzi	computer science	biological networks
CGG ALUMNI		
Dag Ahrén	fungal genomics	metabolic profiling
Benjamin Audit	theoretical physics	genome structure
Alonso Cases	microbiology	gene networks
Richard Coulson	parasitology	transcription factors
Anton Enright	bioinformatics	sequence clustering
Ioannis Iliopoulos	fly genetics	text mining
Paul Janssen	microbiology	genome annotation
David Kreil	bioinformatics	compositional bias
Victor Kunin	bioinformatics	genome evolution
Emmanuel Levy	bioinformatics	error propagation
Núria López	medical genetics	human disease
José (Zé) Pereira	cell biology	protein interactions
José Peregrin	biochemistry	phylogenetic profiling
Mike Smith	applied physics	sequence identifiers

Q&A time…

- Profile entire species with complete genomes, perform comparative analysis?
- Infer metabolic complement for species with incomplete genome information
- Overlay functional information with transcription profiles or inferred networks
- Compare the enzyme complement against entire functional categories
- Use the metabolic complement as a reference for functional module detection
- Reconstruct ancestral states of metabolism via enzymes, or even pathways?