

Scalable Analysis of Microbial Genomes















Morgan Price
Arkin lab
October 2006

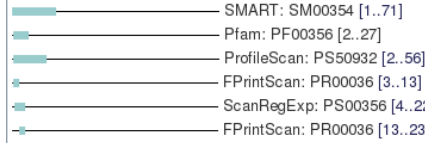

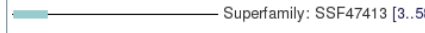



Identifying Similar Sequences

Sequence to sequence (BLAST)

Sequence to family (InterPro, HMMer)

100%		transcriptional repressor for pur regulon, glyA, glnB, prsA, speA	<i>Escherichia coli</i> O157:H7 EDL933	Cart
100%		transcriptional repressor for pur regulon	<i>Escherichia coli</i> O157:H7	Cart
100%		transcriptional repressor for pur regulon, glyA, glnB, prsA, speA	<i>Shigella flexneri</i> 2a str. 2457T	Cart
100%		PurR (see papers)	<i>Shigella boydii</i> Sb227	Cart
100%		DNA-binding transcriptional repressor, hypoxanthine-binding	<i>Escherichia coli</i> W3110	Cart
99.71%		PurR	<i>Shigella dysenteriae</i> Sd197	Cart
99.71%		transcriptional repressor for pur regulon, glyA, glnB, prsA, speA	<i>Shigella sonnei</i> Ss046	Cart
99.71%		purine nucleotide synthesis repressor	<i>Escherichia coli</i> UT189	Cart
99.41%		Purine nucleotide synthesis repressor	<i>Escherichia coli</i> CFT073	Cart
95.89%		transcriptional repressor for pur regulon, glyA, glnB, prsA, speA (GalR/LacI family)	<i>Salmonella typhimurium</i> LT2	Cart
95.89%		transcriptional repressor for pur regulon, glyA, glnB, prsA, speA (GalR/LacI family)	<i>Salmonella enterica</i> Choleraesuis	Cart
95.6%		purine nucleotide synthesis repressor	<i>Salmonella enterica</i> , Typhi	Cart
95.6%		purine nucleotide synthesis repressor	<i>Salmonella enterica</i> , Typhi Ty2	Cart
95.6%		purine nucleotide synthesis repressor	<i>Salmonella enterica</i> Paratyphi A	Cart

VIMSS776892: Putative ribose operon repressor, rbsR (NCBI), 333 a.a. [Photobacterium profundum SS9]	
IPR000843 IPR type: Domain Bacterial regulatory protein, LacI	 <ul style="list-style-type: none"> SMART: SM00354 [1..71] Pfam: PF00356 [2..27] ProfileScan: PS50932 [2..56] FPrintScan: PR00036 [3..13] ScanRegExp: PS00356 [4..22] FPrintScan: PR00036 [13..23]
IPR01761 IPR type: Domain Periplasmic binding protein/LacI transcriptional regulator	 <ul style="list-style-type: none"> Pfam: PF00532 [59..328]
IPR010982 IPR type: Domain Lambda repressor-like, DNA-binding	 <ul style="list-style-type: none"> Superfamily: SSF47413 [3..58]
No IPR id:	
NULL: seg	 <ul style="list-style-type: none"> Seg: seg [10..21]
Info on InterPro	

=> gene's function & evolutionary history

A Rapidly Growing Challenge

- MicrobesOnline has 360 genomes

- ~1,000,000 proteins

- ~300,000,000 amino acids

- Coming soon:

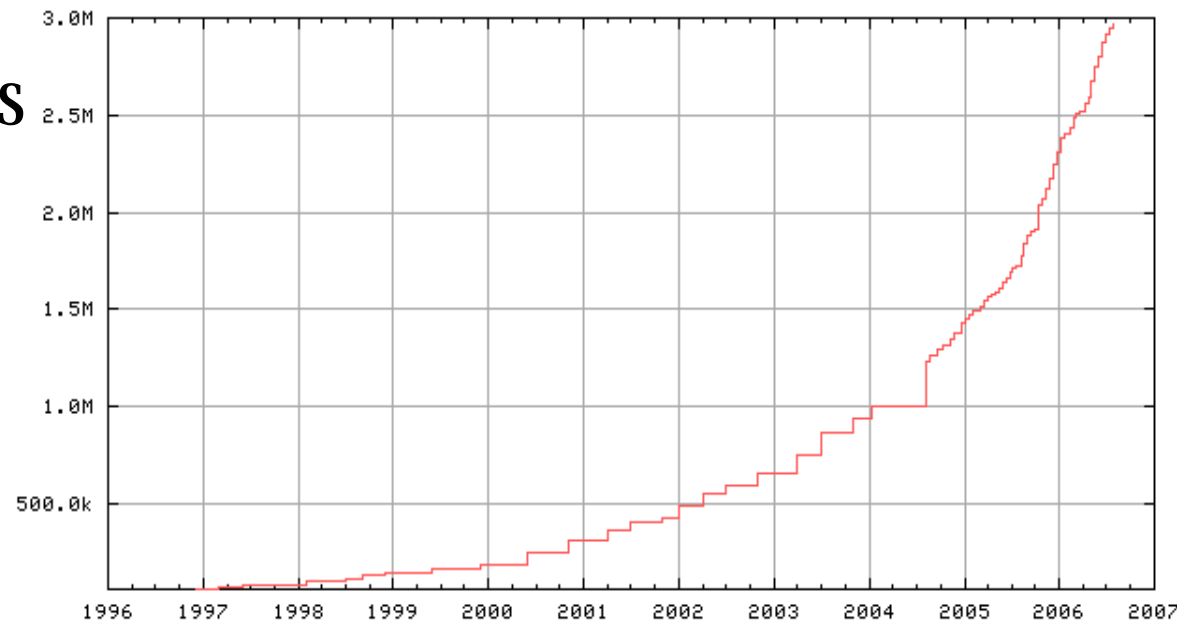
- ~1,000 more genomes

- gigabases from metagenomics

- Computation time already problematic

- delayed updates, supercomputers

UniProt doubles every 2 years



Slow Sequence Analysis Tools

- detect homology with all-vs.-all BLASTp
 - $O(N^2)$ time and storage
 - already 1,000 hits/gene
 - low biological relevance
 - mostly weak hits within known families (~50 bits)
 - no phylogeny
- assign to protein families with HMMs
 - $O(N)$ but slow (>1 CPU-day/genome)

Solutions

- How to view 1,000 homologs
 - Tree-browser
- Fast family assignment with HMMFast
 - combine PSI-BLAST & HMMs
- Fast detection & compact storage of homology
 - rely on known families to avoid most BLASTing

What's Wrong with BLAST hits

transcriptional repressor for pur regulon, glyA, glnB, prsA, speA	<i>Escherichia coli</i> O157:H7 EDL933	Cart
transcriptional repressor for pur regulon	<i>Escherichia coli</i> O157:H7	Cart
transcriptional repressor for pur regulon, glyA, glnB, prsA, speA	<i>Shigella flexneri</i> 2a str. 2457T	Cart
PurR (see papers)	<i>Shigella boydii</i> Sb227	Cart
DNA-binding transcriptional repressor, hypoxanthine-binding	<i>Escherichia coli</i> W3110	Cart
PurR	<i>Shigella dysenteriae</i> Sd197	Cart
transcriptional repressor for pur regulon, glyA, glnB, prsA, speA	<i>Shigella sonnei</i> Ss046	Cart
purine nucleotide synthesis repressor	<i>Escherichia coli</i> UT189	Cart
Purine nucleotide synthesis repressor	<i>Escherichia coli</i> CFT073	Cart
transcriptional repressor for pur regulon, glyA, glnB, prsA, speA (GalR/LacI family)	<i>Salmonella typhimurium</i> LT2	Cart
transcriptional repressor for pur regulon, glyA, glnB, prsA, speA (GalR/LacI family)	<i>Salmonella enterica</i> Choleraesuis	Cart
purine nucleotide synthesis repressor	<i>Salmonella enterica</i> , Typhi	Cart
purine nucleotide synthesis repressor	<i>Salmonella enterica</i> , Typhi Ty2	Cart
purine nucleotide synthesis repressor	<i>Salmonella enterica</i> Paratyphi A	Cart
Transcriptional regulators [Transcription]	<i>Klebsiella pneumoniae</i>	Cart
purine nucleotide synthesis repressor	<i>Yersinia pestis</i> CO92	Cart
transcriptional repressor for pur regulon, glyA, glnB, prsA, speA	<i>Yersinia pestis</i> KIM	Cart
purine nucleotide synthesis repressor	<i>Yersinia pseudotuberculosis</i> IP 32953	Cart
purine nucleotide synthesis repressor	<i>Yersinia pestis</i> biovar Medievalis str. 91001	Cart
transcriptional repressor for pur regulon, glyA, glnB, prsA, speA	<i>Shigella flexneri</i> 2a str. 301	Cart

E. coli purR

- All 20 top hits are from close relatives *Escherichia*, *Shigella*, *Salmonella*, *Klebsiella*, *Yersinia*
- No relatedness
- Little information
(if you know the genus names)

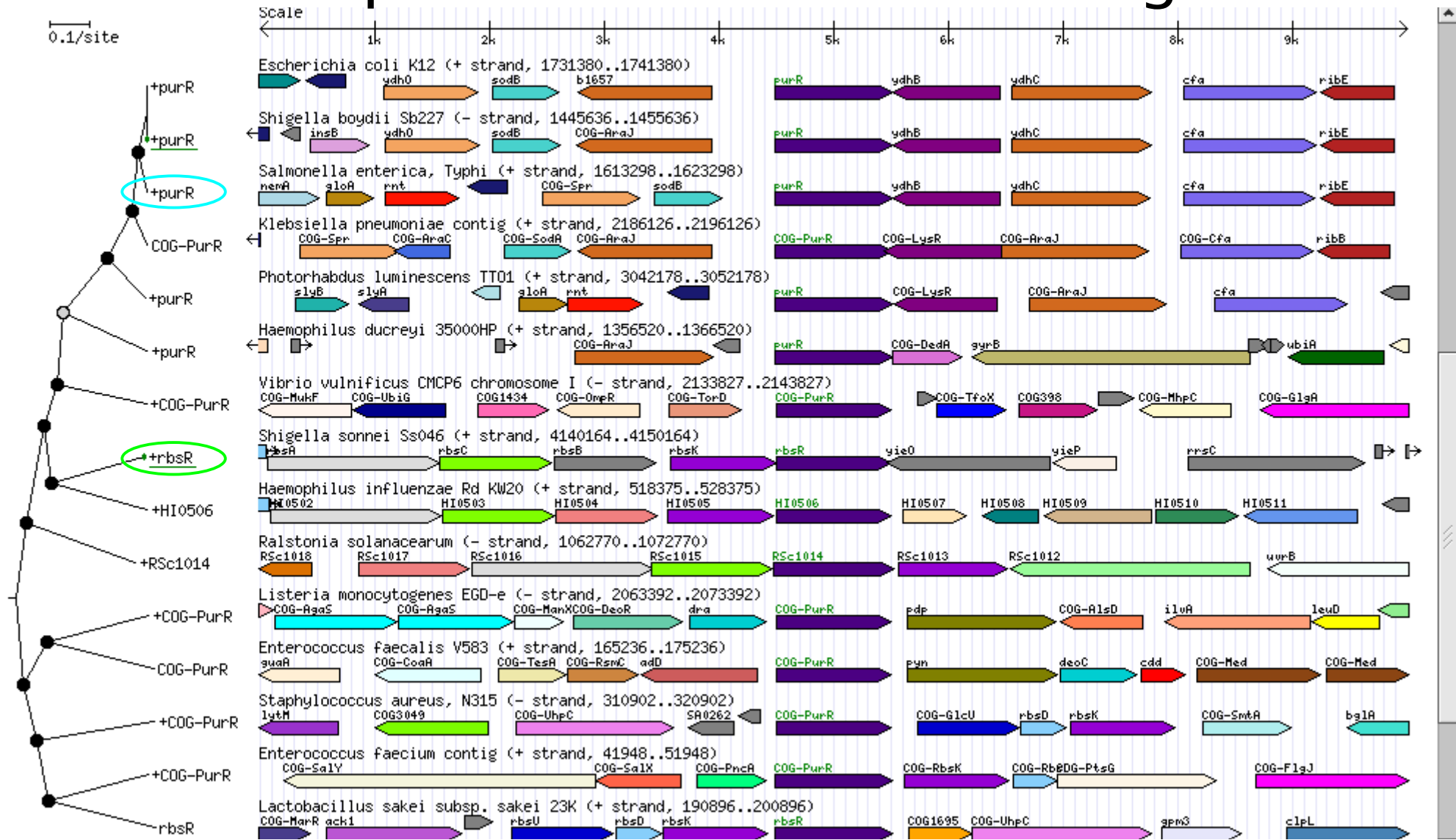
Tree-browser Principles

- Collapse closely related clades so more distant homologies can be seen
 - You can adjust how much to collapse & to show
- Highlight characterized relatives
- Compare domain tree to species tree

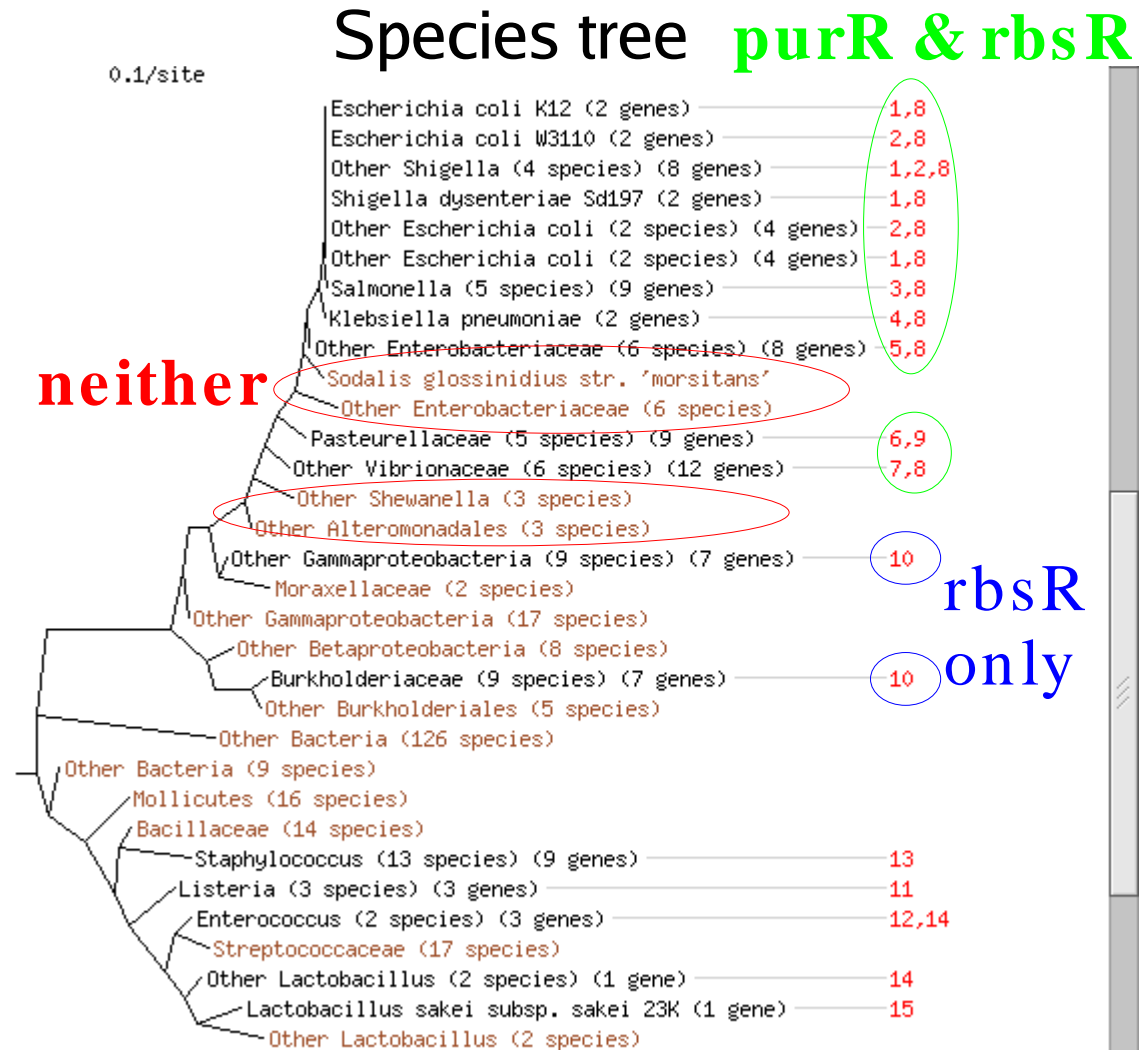
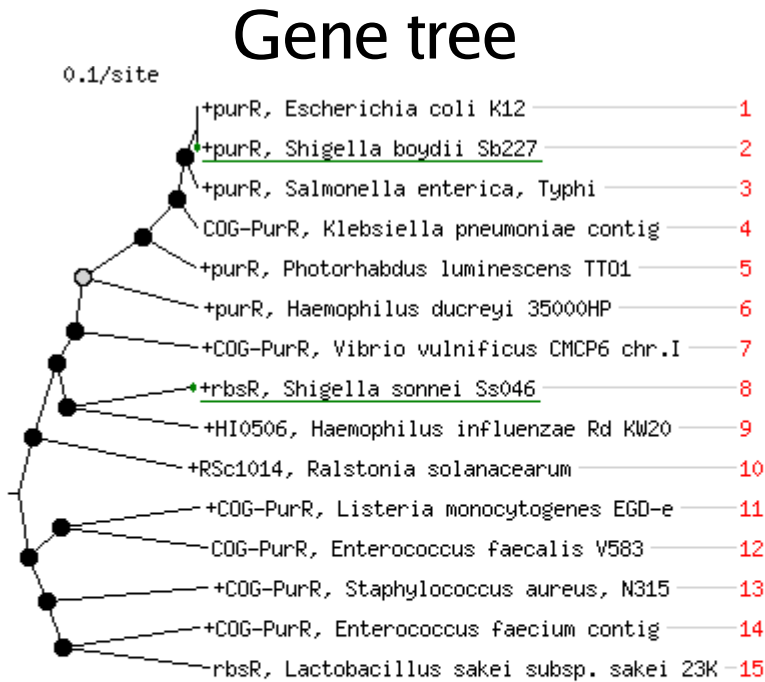
Available at <http://MicrobesOnline.org>

Tree-browser Example

E. coli purR: the 93 closest homologs



Tree-browser Example



Tree-browser Implementation

- Pre-computed tree for every family
 - every Pfam, every COG, and ad-hoc BLAST families for the rest
 - covers >90% of proteins
- Pre-computed species tree
 - from ubiquitous COGs
 - supertree of maximum-likelihood trees

Limitations of Families

Gene tree for VIMSS820723

Help

AM420: hypothetical protein, 487aa

from *Anaplasma marginale* str. St. Maries

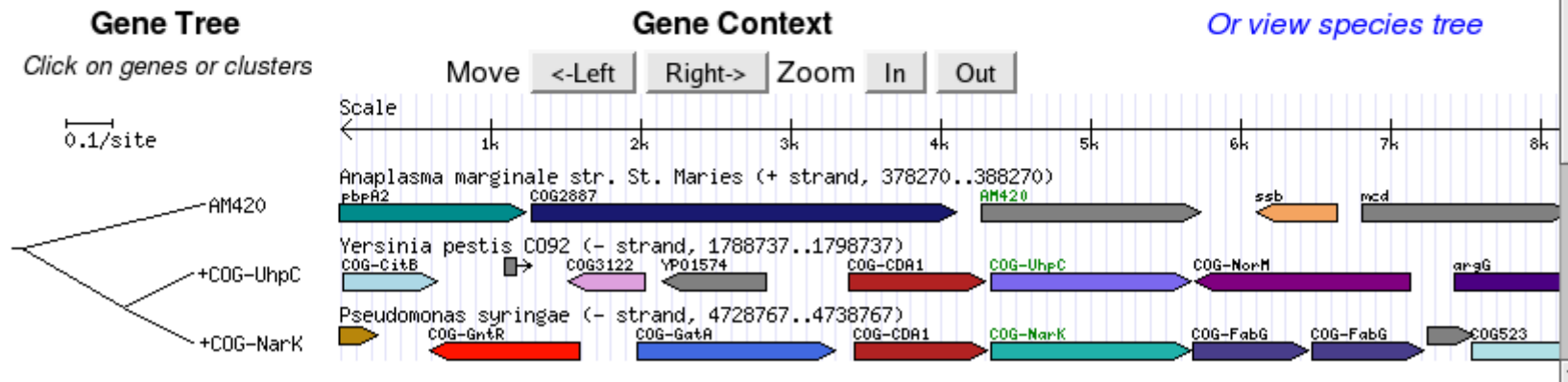
Tree:	Domain used	Range	#Domains	#Shown	Cluster?	#Clusters	Limit
	PF07690 aa 66-432 (1)	aa 66-432 (133)	8855	9	≥ 80% id.	3	≤ 25

Coverage: The top 0 BLAST hits for this gene are in this tree. The best missing hit is VIMSS820718 from *Anaplasma marginale* str. St. Maries (73.64% identity over 440 a.a.). Stop checking BLAST hits

Drawing: Rectangular style? Use branch lengths? Overlapping genes on separate lines?
Color by COG & orthologs of 1st track Exhaustively

Update Reset

the
BLAST
test



Close homologs (74% id.) in different families

Tree-Browser Summary

- Quick access to phylogeny
 - Helps you choose interesting homologs to show
 - Reduces information to manageable size
- Still need to
 - Place genes in known families
 - Group genes not in known families
 - Find the problems with families

HMMFast

- PSI-BLAST is fast, HMMer is accurate
 - HMMer aligns every sequence to every family
 - Sensitive, but slow
 - PSI-BLAST starts with high-scoring hits to profile
 - much faster, but ~15% worse sensitivity/specificity
- Combination
 - Step 1: PSI-BLAST with high sensitivity settings
 - Step 2: run HMMer to remove false positives

HMMFast's Performance

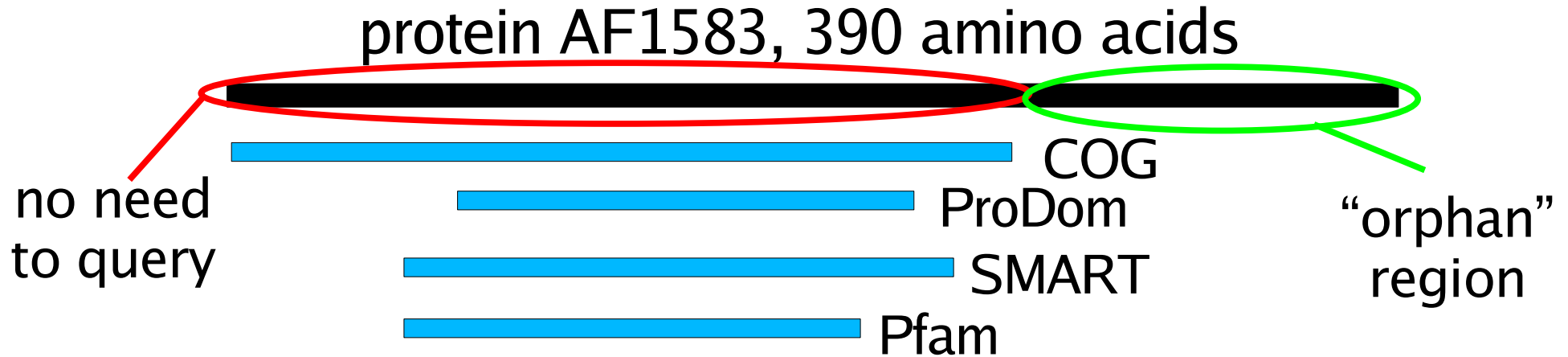
Database	# HMMs	# Test Genomes	HMMer CPU hrs	PSI-BLAST CPU hrs	PSI-BLAST % with hits	Fast hrs (predicted)	% of HMMer Hits Missed
Pfam	,			.	. %	.	.
TIGRfam	,			.	. %	.	.
Supfam	,			.	. %	.	.

- 50-100x faster, >98% sensitivity
- Misses are concentrated in “weak” HMMs
 - half of missing Pfam hits are for families in which a seed sequence has $E > 0.01$ (!)

So Far

- Tree browser
- HMMFast
- Still need BLAST to find ad-hoc domains
- Might need BLAST to find missing family members

Beyond All-vs.-all BLAST



- Use known families
 - cover ~80% of amino acids
- Add ad-hoc “domains”
 - list of hits for an “orphan” region

Reduce Size of Problem

MicrobesOnline Break-down

Millions of a.a.

Total	.
Domains	.
@ %	.
Orphans	.
@ %	.
Spacers	.

- Shrink database by 2.5x
 - Fast $O(N^2)$ clustering within each family (CD-HIT)
 - Also cluster orphan regions
- Shrink query by 7.2x
- $2.5 \times 7.2 = 18x$ faster
 - ~1 CPU-day to preprocess
 - ~3x more by not querying members of new families?

Reconstructing BLAST Scores from Domains

List of domains (including ad-hoc domains)

=> List of genes sharing a domain

=> Rerun BLAST on the small set

Takes ~1 sec./query vs. ~30s on full DB

Performance of Domain-masked BLAST

- 10 test genomes
- Miss 0.24% of homologous pairs of genes
 - only 0.01% of hits of 70 bits or better
 - worst miss is 110 bits (37% id., 182 a.a.)
- Only 0.03% of genes become orphans
- Only 25 domains/gene
 - almost all known families

Not Quite $O(N)$...

$O(N^2)$ Steps Left

- CD-HIT within each family, orphans (fast)
- Domain-masked BLAST
 - Grows slower than N^2 , because of CD-HIT
 - Shrink as families improve?
 - Skip for largest datasets (metagenomics)
- Multiple sequence alignment can be $O(N)$
- Phylogenetic trees...

Remaining Challenges

- Scalable phylogenetic trees
 - large families already contain thousands of members
 - existing methods: fast $O(N^3)$ or slow $O(N^2)$
 - no incremental update
- Tree-based orthologs
- Improved domain family curation
 - automatic splitting
 - build PSI-BLAST models of larger ad-hoc families

Conclusions

Ready for 1,000 genomes

- Tree-browser summarizes a gene's history
 - Available at MicrobesOnline.org
- Fast family assignment, pairwise homology
 - HMMFast going into the analysis pipeline (Keith)

