MAGIC: Integrative and Accurate Comparative Genome Mapping

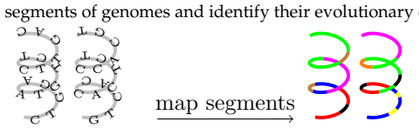
FIRAS SWIDAN[a], EDUARDO P. C. ROCHA[b,c], MICHAEL SHMOISH[a], AND RON Y. PINTER[a]

[a]Department of Computer Science, Technion – Israel Institute of Technology, Haifa 32000, Israel.
[b]Atelier de Bioinformatique, Univ. Paris VI, 12 r Cuvier, 75005 Paris.
[c]Unité GGB, Institut Pasteur, 28 rue Dr Roux, 75015 Paris, France.

1 Introduction

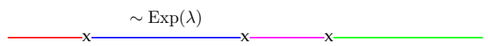
1.1 Comparative Genome Mapping [1]

Map segments of genomes and identify their evolutionary origin.



1.2 The Nadeau-Taylor (NT) Model [2]

Occurrences of **breakpoints** in genomes is modeled by a Poisson process.



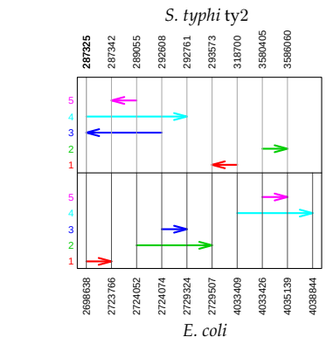
1.3 Questions

- What are the biological events complicating the mapping problem?
- What are the major forces shaping prokaryotic genomes?
- Does the NT model apply to prokaryotic genomes?
- Which forces have likely turned *E. coli* into pathogenic *Shigella*?
- How does mapping differ from alignment?
- How does MAGIC improve on previous work?

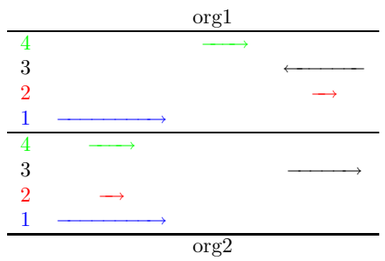
2 Biological Events Complicating the Mapping

- Selfish DNA (incl. Phages)
- Horizontal gene transfer
- Rearrangements
- Duplications (non-selfish DNA)
- Nuisance cross overlaps
- Point mutations

2.1 Rearrangement-Free Segments (RFs), Duplications, and Nuisances



Which hit is consecutive to hit number 1?. These two consecutive hits constitute a **rearrangement-free segment (RF)**.

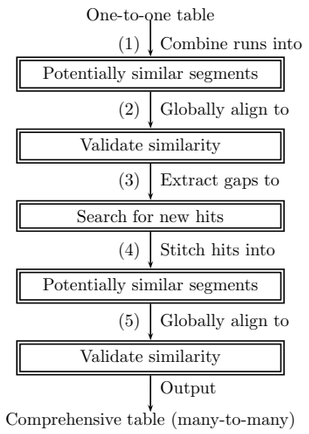


Example of a nuisance cross overlap, an ortholog, and a paralog.

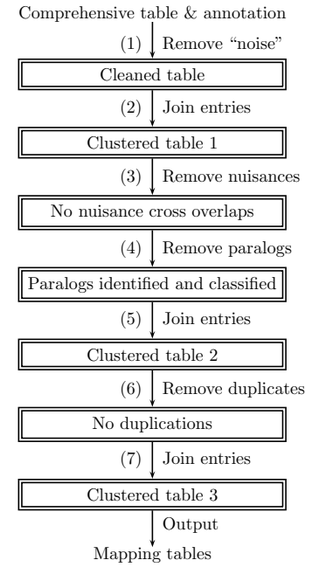
3 MAGIC

MAGIC constitutes of two phases: **preprocessing** and **mapping**.

3.1 The Preprocessing Phase



3.2 The Mapping Phase



4 Results

4.1 Running MAGIC on 10 Pairs of Prokaryotes

Organism	Size	NT	RF(#/cov.)	Orth.	+Par.	+Tr.	+Pro.	Id.
B. aphidicola aps	640681	0.82	1/0.93	0.91	0.91	0.91	0.91	0.75
B. aphidicola sg	641454	0.82	1/0.93	0.90	0.90	0.90	0.90	0.75
E. coli mg1655	4639675	0.050	37/0.94	0.79	0.79	0.82	0.85	0.97
S. flexneri 2457t	4599354	0.052	37/0.93	0.80	0.80	0.88	0.92	0.98
L. monocytogenes	2944528	0.060	11/0.95	0.85	0.85	0.85	0.87	0.87
L. innocua	3011208	0.059	11/0.94	0.83	0.83	0.83	0.90	0.87
P. abyssi	1765118	1.7e-08	99/0.88	0.75	0.75	0.76	0.76	0.73
P. horikoshii	1738505	1.2e-08	99/0.85	0.76	0.76	0.76	0.76	0.73
S. pyogenes m18	1895017	0.46	4/0.96	0.79	0.79	0.79	0.90	0.98
S. pyogenes ssi1	1894275	0.52	4/0.94	0.79	0.79	0.79	0.97	0.98
*B. bronchiseptica**	5339179	0.96	149/0.75	0.68	0.71	0.71	0.71	0.98
*B. pertussis**	4086189	0.95	149/0.92	0.88	0.88	0.91	0.91	0.98
H. pylori	1667867	0.00036	29/0.96	0.92	0.93	0.94	0.94	0.93
H. pylori j99	1643831	0.00038	29/0.96	0.94	0.94	0.94	0.94	0.93
N. meningitidis a	2184406	0.0130	31/0.93	0.90	0.91	0.92	0.95	0.96
N. meningitidis b	2272351	0.0081	31/0.91	0.88	0.89	0.90	0.92	0.96
S. typhi ty2	4791961	0.15	18/0.93	0.85	0.86	0.87	0.90	0.98
S. typhimurium	4857432	0.15	18/0.96	0.84	0.84	0.85	0.90	0.98
Y. pestis co92	4653728	0.66	32/0.97	0.90	0.90	0.94	0.96	0.98
*Y. pseudotuber.**	4744671	0.78	32/0.98	0.89	0.89	0.90	0.90	0.98

Size: genome size. NT: the p-value resulting from the Kolmogorov-Smirnov test. RF(#/cov.): #: number of RF segments. cov.: coverage of RF segments. Orth.: coverage of orthologous segments. +Par.: adding paralogs to the previous column. +Tr.: adding transposable elements to the previous column. +Pro.: adding prophages as well as phagic elements to the previous column. Id.: mean identity of all orthologs. * indicates that no prophage annotation was available for those species.

4.2 Comparing MAGIC to MauVe [3]: Coverage

Organism	Diff. in RFs		Diff. in Orth.		Diff. in OS	
	V-G	G-V	V-G	G-V	V-G	G-V
B. aphidicola aps	0.067	0.0051	0.078	0.027	0.076	0.028
B. aphidicola sg	0.068	0.0051	0.078	0.027	0.076	0.027
E. coli mg1655	0.016	0.041	0.055	0.0044	0.031	0.042
S. flexneri 2457t	0.019	0.020	0.055	0.0032	0.033	0.100
L. monocytogenes	0.047	0.011	0.059	0.0068	0.043	0.012
L. innocua	0.049	0.067	0.057	0.0087	0.038	0.064
P. abyssi	0.026	0.53	0.018	0.50	0.014	0.50
P. horikoshii	0.013	0.56	0.014	0.50	0.014	0.51
S. pyogenes m18	0.022	0.55	0.047	0.44	0.017	0.52
S. pyogenes ssi1	0.035	0.55	0.046	0.43	0.016	0.59
B. bronchiseptica	0.030	0.0250	0.029	0.0110	0.027	0.042
B. pertussis	0.036	0.0074	0.036	0.0044	0.031	0.032
H. pylori	0.0089	0.13	0.012	0.14	0.0096	0.16
H. pylori j99	0.0085	0.12	0.012	0.14	0.0098	0.15
N. meningitidis a	0.036	0.044	0.041	0.045	0.025	0.083
N. meningitidis b	0.049	0.059	0.042	0.060	0.024	0.083
S. typhi ty2	0.040	0.039	0.061	0.0091	0.044	0.041
S. typhimurium	0.036	0.053	0.060	0.0099	0.043	0.046
Y. pestis	0.020	0.017	0.045	0.0043	0.027	0.050
Y. pseudotuber.	0.018	0.040	0.044	0.0083	0.039	0.017

Diff. in RFs coverage: Difference between Mauve's locally collinear blocks (LCBs) coverage and MAGIC's RF coverage. Diff. in Orth.b: Difference between Mauve's Backbone coverage and MAGIC's Orthologs coverage. Diff. in OS (Orthologs & Selfish): Difference between Mauve's Backbone coverage and MAGIC's +Pro. column.

4.3 Comparing MAGIC to Mauve: Mapping

Pair	Transposable			Prophages			Nuisances			Paralogs			Final conflicts			Total		
	#	ℓ_1	ℓ_2	#	ℓ_1	ℓ_2	#	ℓ_1	ℓ_2	#	ℓ_1	ℓ_2	#	ℓ_1	ℓ_2	#	ℓ_1	ℓ_2
B. aphidicola {aps, sg}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E. coli mg1655, *S. flexneri* 2457t	36	32320	31966	44	129024	128641	1	142	143	5	1005	767	0	0	0	86	162491	161517
L. {*monocytogenes* egd-e, *innocua*}	0	0	0	43	77396	77290	0	0	0	6	890	1076	1	54	54	50	78340	78420
P. {*abyssi*, *horikoshii*}	0	0	0	0	0	0	5	1115	1115	10	5580	5345	6	3067	2924	21	9762	9384
S. pyogenes {m18, ssi1}	0	0	0	63	35511	35594	1	619	619	2	240	238	0	0	0	66	36370	36451
B. {*bronchiseptica*, *pertussis*}	0	0	0	0	0	0	10	32528	32432	6	2352	1998	8	29415	29166	24	64295	63596
H. pylori {j99}	2	1380	1409	0	0	0	19	24872	24533	2	1140	1140	8	26370	26425	31	53762	53507
N. meningitidis {a, b}	51	12720	12674	13	20167	20067	33	32078	32089	39	64031	63641	94	422955	423515	230	551951	551986
S. {*typhi* ty2, *typhimurium*}	19	11911	12260	34	82543	82830	6	10522	10581	14	3303	2950	0	0	0	73	108279	108621
Y. {*pestis*, *pseudotuberculosis*}	23	21703	22721	21	109399	109775	5	1003	1003	10	4566	4365	22	196986	196821	81	333657	334685

Running MAGIC on Mauve's backbones and classifying the backbones into 5 categories. Transposable, Prophages, Nuisances, Paralogs, and Final Conflicts: Backbones that were identified (by MAGIC) into the corresponding categories. Total: summing over the five categories. #: number of backbones. ℓ_1/ℓ_2: length of the backbones in the first/second organism. The classification is based on an intersection greater than 90%.

5 Discussion

5.1 MAGIC's Results (Section 4.1)

- Indels seem to be the major force shaping prokaryotic evolution: Large deletions contribute up to 30% in *B. pertussis* [4]. Transposable elements contribute up to 8% in *S. flexneri* 2457t, and prophages contribute up to 18% in *S. pyogenes* ssi1.
- **Duplications** (not resulting from selfish DNA) **contribute less than** 1%, except in *B. bronchiseptica* (3%).
- MAGIC's results fit the NT model predictions in most cases.

5.1 Comparison with Mauve (Sections 4.2 and 4.3)

- Mauve results sometimes in poor coverages (when run with default parameters). For example, the difference between the coverage of MAGIC and of Mauve reaches more than 50% in *P. horikoshii*.
- Even when Mauve's and MAGIC's coverages are similar, there are usually significant differences in the mappings. For example, this difference reaches up to 25% in *N. meningitidis* a.
- Some differences result from MAGIC taking known selfish DNA into consideration (*e.g.*, in *S. flexneri* 2457t). Other result from a different handling of duplications (*e.g.*, *N. meningitidis* a).
- Mauve's results reject the NT model in almost all pairs (except in *buchnera*, data not shown).
- Mauve is faster than MAGIC. On the pair *S. flexneri* 2457t vs. *S. typhi* ty2, Mauve required less than 17 min of CPU time, while MAGIC consumed 35 min (data not shown).
- Mauve is applicable for multiple comparisons, MAGIC only for pairwise (in the meanwhile).

6 Conclusions

- Mapping is harder than alignment.
- Measuring the differences between mappings based on nucleotide percentages is not optimal: The nucleotide difference between MAGIC's and Mauve's mappings on the pair *E. coli* mg1655 vs. *S. flexneri* 2457t is less than 4%. Yet the fragmentation resulting from the two mappings is significantly different.
- Validating the NT model is a very delicate matter. Inadequate identification of selfish DNA or handling of duplications can cause a significant bias.

References

[1] Swidan, F., Rocha, E., Shmoish, M., and Pinter, R. Y., An integrative method for accurate genome mapping, (2005).

[2] Nadeau, J. H. and Taylor, B. A., Lengths of chromosomal segments conserved since divergence of man and mouse, Proc. Natl. Acad. Sci. USA **81** (1984) 814.

[3] Darling, A. C., Mau, B., Blattner, F. R., and Perna, N. T., Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements, Genome Res. **14** (2004) 1394.

[4] Parkhill, J. et al., Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*, Nat. Genet. **35** (2003) 32.