Universidade de São Paulo
Instituto de Química

Comparação de sequências

aula 4

João Carlos Setubal

2024

Alinhamento múltiplo de sequências longas

- O programa MAUVE
- Darling AC, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 2004 Jul;14(7):1394-403.

How MAUVE works

- Seed-and-extend hashing
- Seeds/anchors: Maximal Multiple Unique Matches of minimum length *k*
- Result: Local Collinear Blocks (LCBs)
- O(G^2n + Gn log Gn), G = # genomes, n = average genome length

Alignment algorithm

1. Find Multi-MUMs
2. Use the multi-MUMs to calculate a phylogenetic guide tree
3. Find LCBs (subset of multi-MUMs; filter out spurious matches; requires *minimum weight*)
4. Recursive anchoring to identify additional anchors (extension of LCBs)
5. Progressive alignment (CLUSTALW) using guide tree

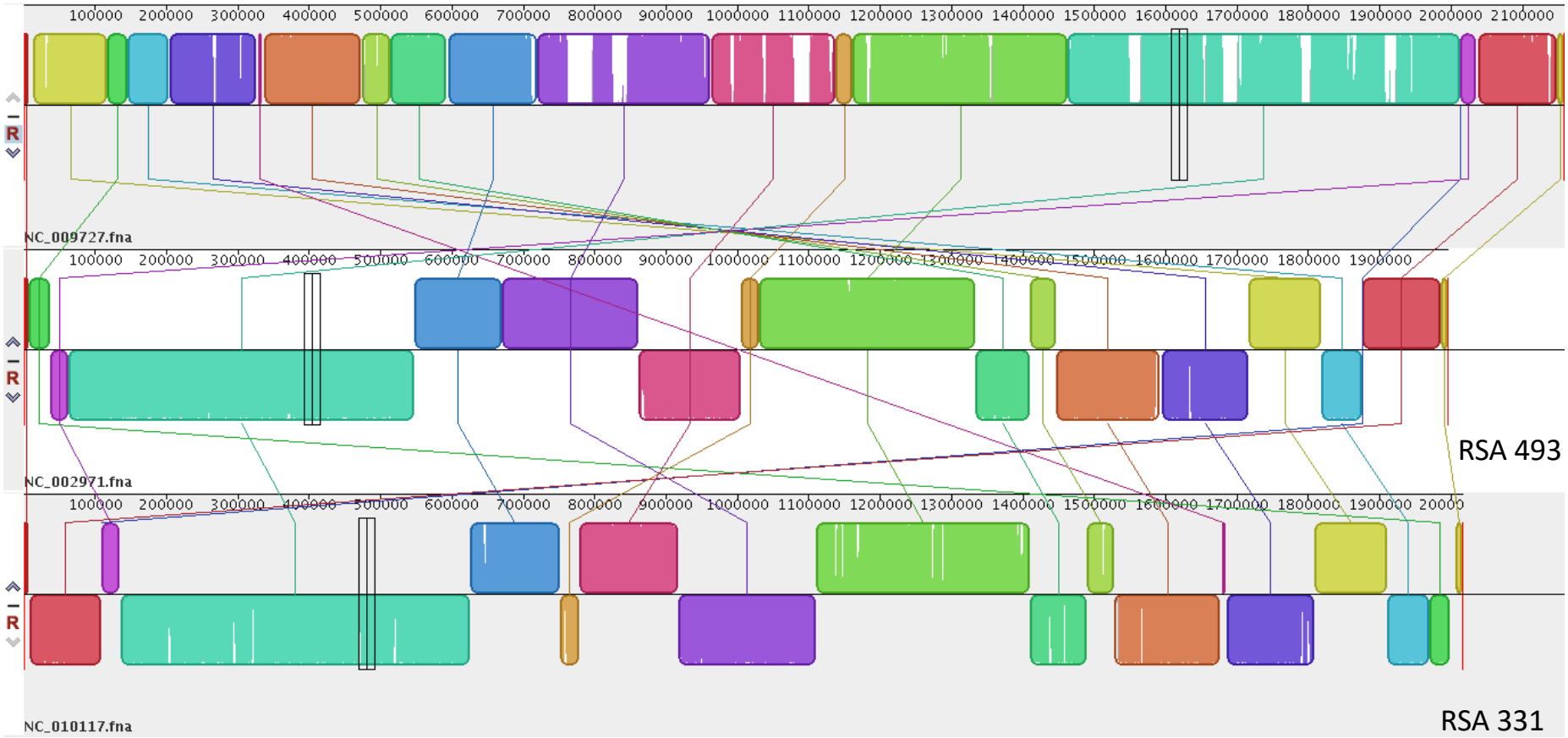
Exemplos de alinhamentos obtidos com MAUVE

- As cores indicam os LCBs encontrados

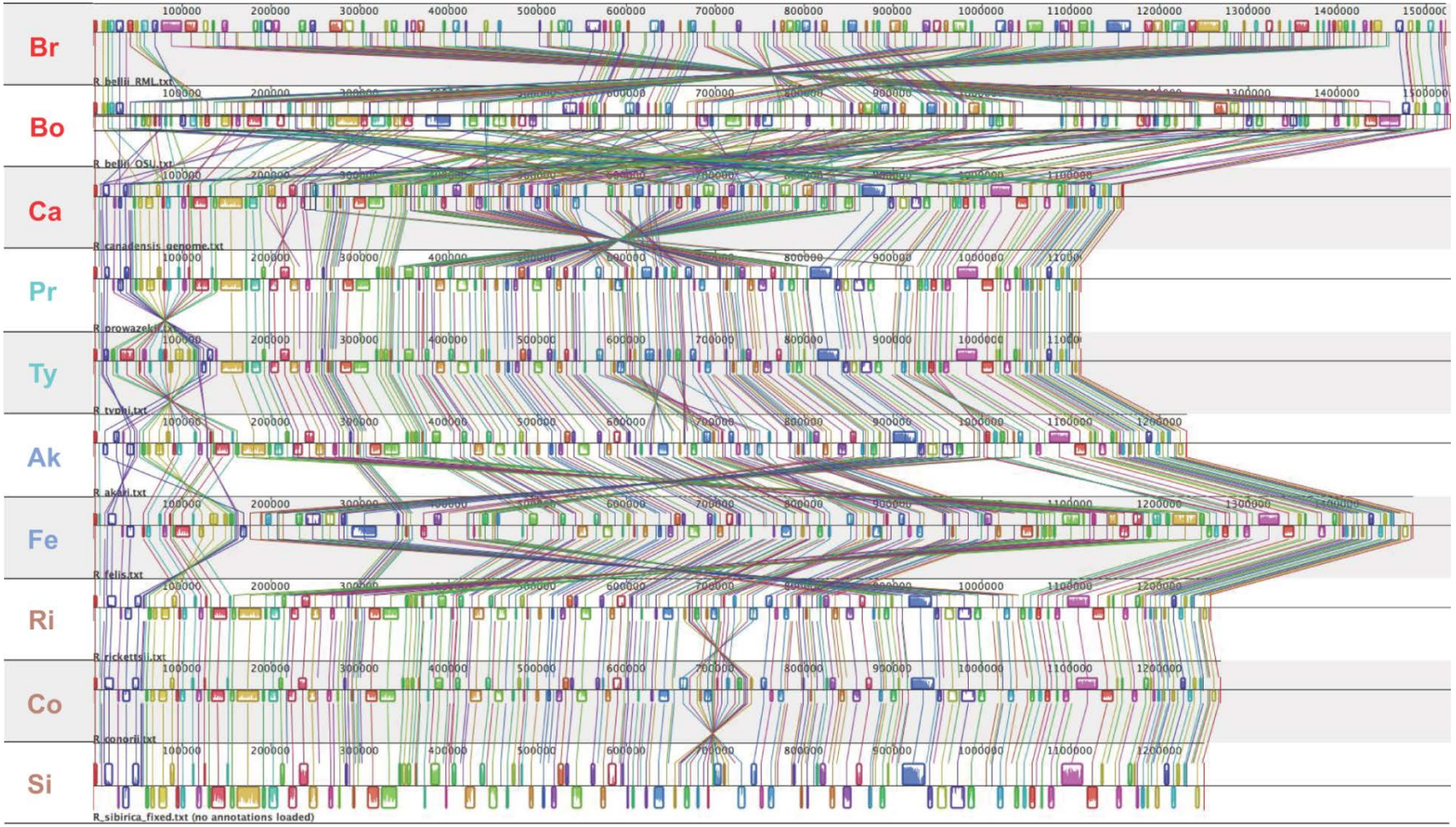
Brucella: Main chromosome alignment

Brucella: Chromosome 2 alignment

Coxiella: Chromosome alignment



Rickettsia


Br
Bo
Ca
Pr
Ty
Ak
Fe
Ri
Co
Si
R_bellii_RML.txt
R_bellii_OSU.txt
R_canadensis_genome.txt
R_prowazekii.txt
R_typhi.txt
R_akari.txt
R_felis.txt
R_rickettsii.txt
R_conorii.txt
R_sibirica_fixed.txt (no annotations loaded)

Exercício

- Usando o serviço de alinhamento de genomas da plataforma BV-BRC.org, compare 2 ou mais cromossomos de bactérias
- https://docs.patricbrc.org/user_guides/services/genome_alignment_service.html
- Esse serviço roda MAUVE

Sumário de comparação de sequências

	Sequências curtas	**Sequências longas**
2-a-2	Prog. Dinâmica	Mummer
2-a-2 muitas vezes	BLAST, usearch, diamond	Mummer
múltiplo	Muscle, MAFFT	Mauve, MUGSY

Distância genômica

- Ao comparar genomas, muitas vezes é útil poder expressar essa comparação por meio de um único número
  - Quando se comparam pares de replicons
    - Que podem ser “replicons” concatenados
- Distância pode ser entendida como o inverso da similaridade

Distância e similaridade

- São conceitos muito parecidos
- Em particular *distância de edição*
- Como transformar sequência *s* em sequência *t*
- Operações
  - Substituição do caracter *a* por *b* (custo = 1)
  - Inserção ou Remoção de um caracter (custo = 2)
- O algoritmo de PD já visto resolve esse problema (com pequenas modificações)

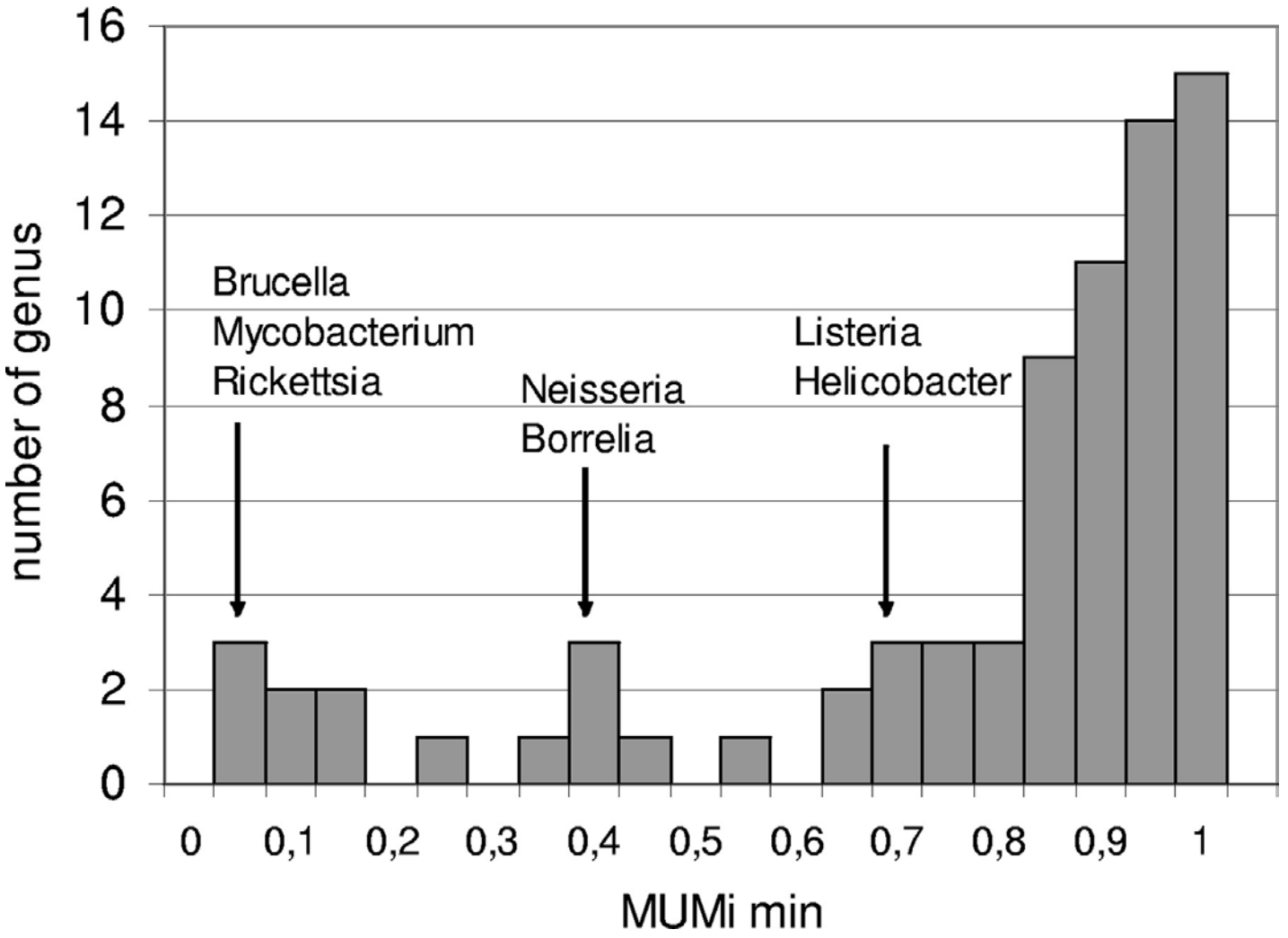
Uma fórmula de distância genômica

- MUMi = MUM index
- Baseado em MUMmer
- **Deloger et al. 2009**
- MUMi = $1 - L_{mum}/L_{av}$
- L_{mum} = soma dos comprimentos de todos os MUMs que não tem sobreposição
- L_{av} = comprimento médio dos 2 genomas sendo comparados
- Para obter MUMi, basta rodar MUMmer com um script perl desses autores

O que significa o índice MUMi

- Identical sequences: zero
- Totally different sequences: 1
- The boundary between **genus** and **species** is around 0.8

Distribution of all minimal MUMi values per genus.



Marc Deloger et al. J. Bacteriol. 2009;191:91-99

Journal of Bacteriology

Conclusão

- Não dá para comparar distâncias MUMi entre diferentes gêneros; apenas entre diferentes espécies do mesmo gênero
- ou entre diferentes cepas da mesma espécie

Uma matriz de distâncias genômicas em Brucella

Strain	MUMi value[b]					
	83–13	BO2	NF 2653	BO1	*B. suis*	*B. microti*
B. neotomae 5K33	0.145	0.168	0.146	0.168	0.017	0.022
Brucella sp. strain 83–13		0.172	0.009	0.175	0.147	0.150
Brucella sp. strain BO2			0.168	0.107	0.169	0.169
Brucella sp. strain NF 2653				0.172	0.147	0.146
B. inopinata BO1					0.169	0.167
B. suis biovar 3 686						0.020

↵*a* Only the sequence of *B. microti* was complete, and its two chromosome sequences were concatenated. Concatenation was done by inserting a string with 100 nucleotides between contigs.

↵*b* The minimum MUMi value is 0, and the maximum MUMi value is 1.

Valores MUMi em Xanthomonas

XacA628	0.4017	0.0176	0.0172	0.0090	0.0193		0.0084	0.0083	0.0069	0.0606	0.0523	0.0558	0.0536	0.0536	0.0535	0.0536	0.0482	0.0538	0.0169	0.0169	0.0149	0.015(
XacA636	0.4030	0.0177	0.0182	0.0088	0.0238	0.0084		0.0085	0.0116	0.0624	0.0550	0.0584	0.0560	0.0560	0.0560	0.0560	0.0511	0.0561	0.0179	0.0180	0.0169	0.016
XacA654	0.4033	0.0131	0.0136	0.0036	0.0251	0.0083	0.0085		0.0119	0.0625	0.0557	0.0592	0.0568	0.0568	0.0568	0.0568	0.0512	0.0569	0.0133	0.0133	0.0180	0.018
XacA828	0.4040	0.0208	0.0211	0.0123	0.0178	0.0069	0.0116	0.0119		0.0646	0.0558	0.0598	0.0565	0.0565	0.0565	0.0565	0.0518	0.0566	0.0210	0.0210	0.0181	0.018
XacAS270	0.4073	0.0656	0.0655	0.0629	0.0707	0.0606	0.0624	0.0625	0.0646		0.0131	0.0067	0.0603	0.0600	0.0603	0.0600	0.0352	0.0601	0.0654	0.0654	0.0639	0.063
XacAS8	0.4031	0.0590	0.0588	0.0558	0.0633	0.0523	0.0550	0.0557	0.0558	0.0131		0.0085	0.0541	0.0538	0.0541	0.0538	0.0291	0.0539	0.0588	0.0588	0.0561	0.056
XacAS9	0.4045	0.0621	0.0619	0.0593	0.0661	0.0558	0.0584	0.0592	0.0598	0.0067	0.0085		0.0558	0.0555	0.0558	0.0555	0.0304	0.0556	0.0619	0.0619	0.0578	0.057
XacAW13	0.4074	0.0584	0.0581	0.0567	0.0637	0.0536	0.0560	0.0568	0.0565	0.0603	0.0541	0.0558		0.0001	0.0001	0.0001	0.0611	0.0004	0.0578	0.0578	0.0554	0.055
XacAW14	0.4074	0.0581	0.0578	0.0568	0.0633	0.0536	0.0560	0.0568	0.0565	0.0600	0.0538	0.0555	0.0001		0.0001	0.0001	0.0611	0.0004	0.0575	0.0575	0.0551	0.055
XacAW15	0.4074	0.0584	0.0581	0.0567	0.0637	0.0535	0.0560	0.0568	0.0565	0.0603	0.0541	0.0558	0.0001	0.0001		0.0001	0.0610	0.0004	0.0578	0.0578	0.0554	0.055
XacAW16	0.4074	0.0581	0.0578	0.0568	0.0633	0.0536	0.0560	0.0568	0.0565	0.0600	0.0538	0.0555	0.0001	0.0001	0.0001		0.0611	0.0004	0.0575	0.0575	0.0551	0.055
XacAs1682	0.4007	0.0599	0.0597	0.0518	0.0643	0.0482	0.0511	0.0512	0.0518	0.0352	0.0291	0.0304	0.0611	0.0611	0.0610	0.0611		0.0613	0.0594	0.0594	0.0574	0.057
XacAw12879	0.4074	0.0583	0.0579	0.0568	0.0636	0.0538	0.0561	0.0569	0.0566	0.0601	0.0539	0.0556	0.0004	0.0004	0.0004	0.0004	0.0613		0.0576	0.0577	0.0553	0.055
XacBL18	0.4094	0.0007	0.0003	0.0125	0.0222	0.0169	0.0179	0.0133	0.0210	0.0654	0.0588	0.0619	0.0578	0.0575	0.0578	0.0575	0.0594	0.0576		0.0000	0.0063	0.006
XacFB19	0.4094	0.0007	0.0003	0.0125	0.0222	0.0169	0.0180	0.0133	0.0210	0.0654	0.0588	0.0619	0.0578	0.0575	0.0578	0.0575	0.0594	0.0577	0.0000		0.0063	0.006
XacGD2	0.4077	0.0057	0.0063	0.0171	0.0194	0.0149	0.0169	0.0180	0.0181	0.0639	0.0561	0.0578	0.0554	0.0551	0.0554	0.0551	0.0574	0.0553	0.0063	0.0063		0.000
XacGD3	0.4077	0.0058	0.0063	0.0171	0.0194	0.0150	0.0169	0.0181	0.0182	0.0639	0.0561	0.0579	0.0557	0.0553	0.0557	0.0553	0.0575	0.0555	0.0064	0.0064	0.0005	
XacJX4	0.4077	0.0065	0.0070	0.0168	0.0196	0.0146	0.0173	0.0177	0.0178	0.0641	0.0563	0.0579	0.0557	0.0554	0.0557	0.0554	0.0572	0.0556	0.0068	0.0071	0.0024	0.002
XacJX5	0.4077	0.0065	0.0070	0.0167	0.0196	0.0146	0.0172	0.0177	0.0179	0.0640	0.0561	0.0578	0.0556	0.0553	0.0556	0.0553	0.0571	0.0555	0.0071	0.0071	0.0024	0.002
XacLMG941	0.4125	0.1430	0.1429	0.1354	0.1467	0.1324	0.1345	0.1348	0.1348	0.1380	0.1364	0.1365	0.1535	0.1535	0.1535	0.1535	0.1309	0.1537	0.1427	0.1427	0.1388	0.138
XacMF20	0.4094	0.0007	0.0003	0.0125	0.0222	0.0170	0.0180	0.0133	0.0210	0.0652	0.0587	0.0617	0.0578	0.0575	0.0578	0.0575	0.0595	0.0577	0.0000	0.0000	0.0063	0.006
XacMN10	0.4076	0.0056	0.0062	0.0168	0.0195	0.0144	0.0171	0.0176	0.0177	0.0638	0.0560	0.0576	0.0556	0.0553	0.0556	0.0553	0.0570	0.0555	0.0062	0.0062	0.0024	0.002
XacMN11	0.4075	0.0052	0.0058	0.0169	0.0186	0.0146	0.0172	0.0177	0.0182	0.0639	0.0559	0.0577	0.0560	0.0557	0.0560	0.0557	0.0569	0.0559	0.0059	0.0058	0.0026	0.002
XacMN12	0.4076	0.0054	0.0060	0.0166	0.0186	0.0144	0.0169	0.0175	0.0177	0.0638	0.0560	0.0577	0.0556	0.0553	0.0556	0.0553	0.0569	0.0555	0.0060	0.0061	0.0023	0.002
XacNT17	0.4095	0.0005	0.0001	0.0125	0.0221	0.0171	0.0180	0.0135	0.0209	0.0656	0.0589	0.0621	0.0580	0.0577	0.0580	0.0577	0.0596	0.0578	0.0002	0.0002	0.0062	0.006
XacUI6	0.4077	0.0059	0.0065	0.0171	0.0190	0.0149	0.0177	0.0181	0.0182	0.0644	0.0566	0.0583	0.0558	0.0555	0.0558	0.0555	0.0572	0.0557	0.0063	0.0065	0.0025	0.002
XacUI7	0.4076	0.0055	0.0060	0.0173	0.0190	0.0150	0.0177	0.0182	0.0183	0.0644	0.0567	0.0583	0.0557	0.0554	0.0557	0.0554	0.0571	0.0556	0.0061	0.0061	0.0014	0.001
XacX18	0.3933	0.1462	0.1461	0.1379	0.1488	0.1348	0.1368	0.1378	0.1364	0.1522	0.1446	0.1455	0.1543	0.1543	0.1543	0.1543	0.1399	0.1544	0.1459	0.1459	0.1422	0.142
XacX20	0.4086	0.1559	0.1559	0.1476	0.1590	0.1447	0.1469	0.1476	0.1462	0.1417	0.1337	0.1351	0.1471	0.1471	0.1471	0.1471	0.1367	0.1472	0.1557	0.1557	0.1519	0.151
XagCFBP7119	0.4154	0.1611	0.1608	0.1537	0.1650	0.1507	0.1531	0.1535	0.1521	0.1552	0.1469	0.1488	0.1544	0.1544	0.1544	0.1544	0.1453	0.1545	0.1606	0.1606	0.1574	0.157
XalfaCFBP3836	0.1448	0.3896	0.3893	0.3831	0.3930	0.3808	0.3822	0.3826	0.3832	0.3924	0.3882	0.3888	0.3951	0.3952	0.3952	0.3951	0.3839	0.3952	0.3892	0.3893	0.3867	0.386
XalfaF1	0.1413	0.3926	0.3923	0.3876	0.3965	0.3852	0.3866	0.3870	0.3868	0.3951	0.3905	0.3916	0.3980	0.3980	0.3980	0.3980	0.3871	0.3980	0.3921	0.3921	0.3900	0.390
Xalfacm1510	0.1390	0.3895	0.3892	0.3829	0.3935	0.3807	0.3810	0.3824	0.3822	0.3910	0.3864	0.3875	0.3942	0.3942	0.3942	0.3942	0.3822	0.3943	0.3891	0.3891	0.3871	0.387
Xalfacm1637	0.1398	0.3903	0.3901	0.3832	0.3932	0.3814	0.3810	0.3831	0.3823	0.3938	0.3888	0.3901	0.3960	0.3960	0.3960	0.3960	0.3853	0.3961	0.3899	0.3899	0.3874	0.387
XauB11122	0.4191	0.2843	0.2842	0.2767	0.2833	0.2726	0.2749	0.2762	0.2709	0.2837	0.2811	0.2829	0.2919	0.2919	0.2919	0.2919	0.2792	0.2919	0.2841	0.2841	0.2808	0.280
XauB1561	0.4264	0.2948	0.2946	0.2872	0.2984	0.2854	0.2857	0.2874	0.2867	0.2949	0.2925	0.2941	0.3021	0.3021	0.3021	0.3021	0.2902	0.3022	0.2945	0.2945	0.2914	0.291
XauC10535	0.4164	0.2877	0.2876	0.2800	0.2912	0.2777	0.2784	0.2798	0.2798	0.2912	0.2853	0.2876	0.2962	0.2962	0.2962	0.2962	0.2823	0.2962	0.2874	0.2874	0.2842	0.284
XauC1559	0.4412	0.3166	0.3166	0.3101	0.3201	0.3085	0.3079	0.3104	0.3100	0.3216	0.3156	0.3178	0.3255	0.3255	0.3256	0.3255	0.3126	0.3256	0.3164	0.3164	0.3135	0.313
XauC1609	0.4354	0.3036	0.3034	0.2974	0.3070	0.2951	0.2959	0.2972	0.2969	0.3081	0.3021	0.3039	0.3130	0.3130	0.3130	0.3129	0.2991	0.3130	0.3033	0.3033	0.3004	0.300
XauC535	0.4241	0.2953	0.2953	0.2875	0.2989	0.2854	0.2861	0.2872	0.2871	0.2984	0.2919	0.2941	0.3041	0.3041	0.3041	0.3041	0.2890	0.3041	0.2951	0.2951	0.2921	0.292
XauC763	0.4109	0.2807	0.2806	0.2729	0.2843	0.2702	0.2710	0.2722	0.2721	0.2844	0.2783	0.2805	0.2894	0.2894	0.2894	0.2894	0.2750	0.2895	0.2804	0.2804	0.2772	0.277

XacMN12	0.4076	0.0054	0.0060	0.0166	0.0186	0.0144	0.0169	0.0175	0.0177	0.063
XacNT17	0.4095	0.0005	0.0001	0.0125	0.0221	0.0171	0.0180	0.0135	0.0209	0.065
XacUI6	0.4077	0.0059	0.0065	0.0171	0.0190	0.0149	0.0177	0.0181	0.0182	0.064
XacUI7	0.4076	0.0055	0.0060	0.0173	0.0190	0.0150	0.0177	0.0182	0.0183	0.064
XacX18	0.3933	0.1462	0.1461	0.1379	0.1488	0.1348	0.1368	0.1378	0.1364	0.152
XacX20	0.4086	0.1559	0.1559	0.1476	0.1590	0.1447	0.1469	0.1476	0.1462	0.141
XagCFBP7119	0.4154	0.1611	0.1608	0.1537	0.1650	0.1507	0.1531	0.1535	0.1521	0.155
XalfaCFBP3836	0.1448	0.3896	0.3893	0.3831	0.3930	0.3808	0.3822	0.3826	0.3832	0.392
XalfaF1	0.1413	0.3926	0.3923	0.3876	0.3965	0.3852	0.3866	0.3870	0.3868	0.395
Xalfacm1510	0.1390	0.3895	0.3892	0.3829	0.3935	0.3807	0.3810	0.3824	0.3822	0.391
Xalfacm1637	0.1398	0.3903	0.3901	0.3832	0.3932	0.3814	0.3810	0.3831	0.3823	0.393
XauB11122	0.4191	0.2843	0.2842	0.2767	0.2833	0.2726	0.2749	0.2762	0.2709	0.283
XauB1561	0.4264	0.2948	0.2946	0.2872	0.2984	0.2854	0.2857	0.2874	0.2867	0.294
XauC10535	0.4164	0.2877	0.2876	0.2800	0.2912	0.2777	0.2784	0.2798	0.2798	0.291
XauC1559	0.4412	0.3166	0.3166	0.3101	0.3201	0.3085	0.3079	0.3104	0.3100	0.321
XauC1609	0.4354	0.3036	0.3034	0.2974	0.3070	0.2951	0.2959	0.2972	0.2969	0.308
XauC535	0.4241	0.2953	0.2953	0.2875	0.2989	0.2854	0.2861	0.2872	0.2871	0.298
XauC763	0.4199	0.2907	0.2906	0.2829	0.2943	0.2799	0.2810	0.2822	0.2821	0.294

Largest distance: **0.4506** (*X. perforans* 91-118 and *X. fuscans aurantifolii* C 1559)

For comparison: distance between *Stenotrophomonas maltophilia* and Xac 306: **0.912**

ANI

- Average Nucleotide Identity [Goris et al. 2007]
- Baseado em BLAST
- one-way ANI (best hits)
- two-way ANI (reciprocal best hits)
- Typically, 2 species will have ANI ≥ 95%
- < 75% : not to be trusted
- http://enve-omics.ce.gatech.edu/ani/

FastANI

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-07641-9 OPEN

High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries

Chirag Jain[1,2], Luis M. Rodriguez-R[3,4], Adam M. Phillippy[2], Konstantinos T. Konstantinidis[3,4] & Srinivas Aluru[1,5]

FastANI can be downloaded at https://github.com/ParBLiSS/FastANI/releases.

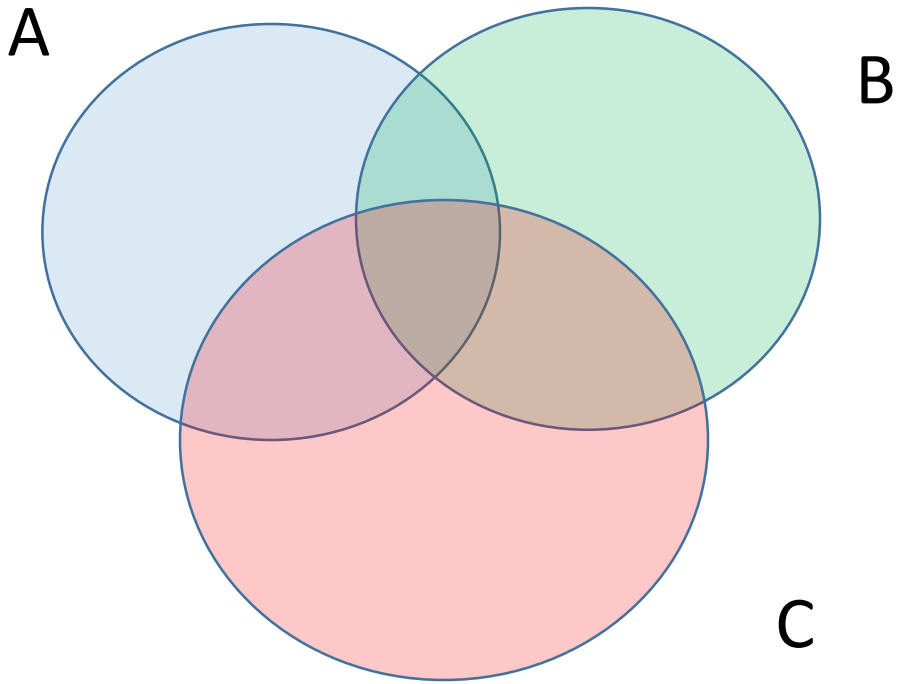
GGDC

- Genome-to-genome distance calculator
- rigorous *in silico* replacement for DNA-DNA hybridization wet-lab experiments
- http://ggdc.dsmz.de
- Meier-Kolthoff, J.P., et al., *Genome sequence-based species delimitation with confidence intervals and improved distance functions.* BMC Bioinformatics, 2013. 14: p. 60

Comparação de conjuntos de genes

- Given a set of genomes, represented by their ‘proteomes’ or sets of protein sequences
- Given homologous relationships (as given for example by orthoMCL, veja adiante)
  - Which genes are shared by genomes *X* and *Y*?
  - Which genes are unique to genome *Z*?
  - Venn or extended Venn diagrams

3-way genome comparison



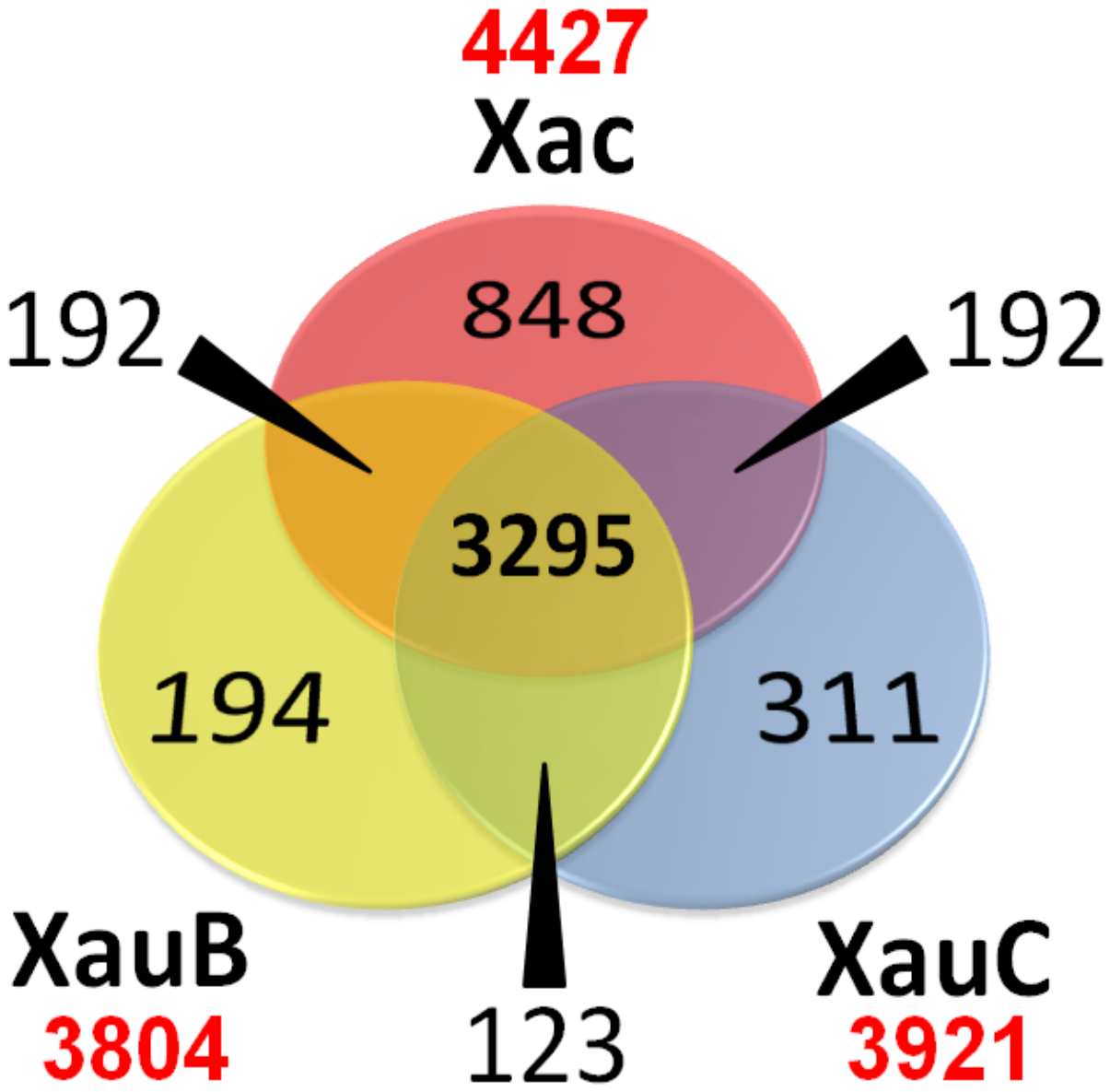
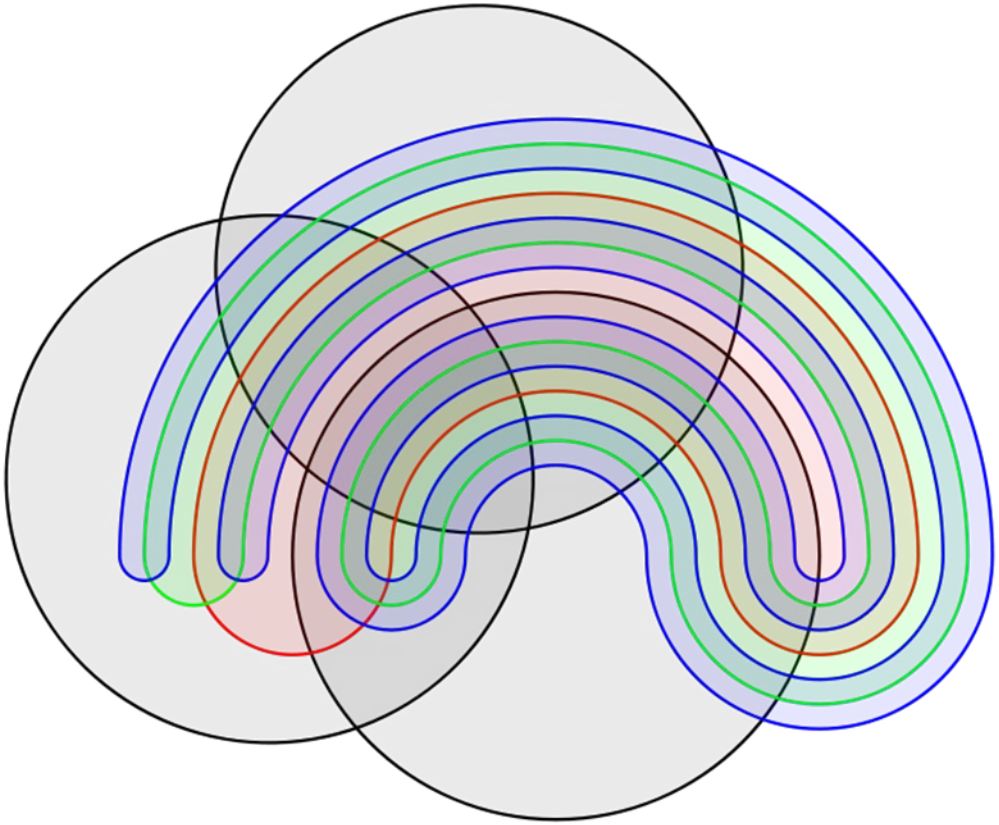

4427
Xac
192
848
192
3295
194
311
XauB
3804
123
XauC
3921

Diagrama de Venn para n = 6



Número de comparações é quadrático em *n*
Número de regiões num diagrama de Venn = 2^n

Source: wikipedia

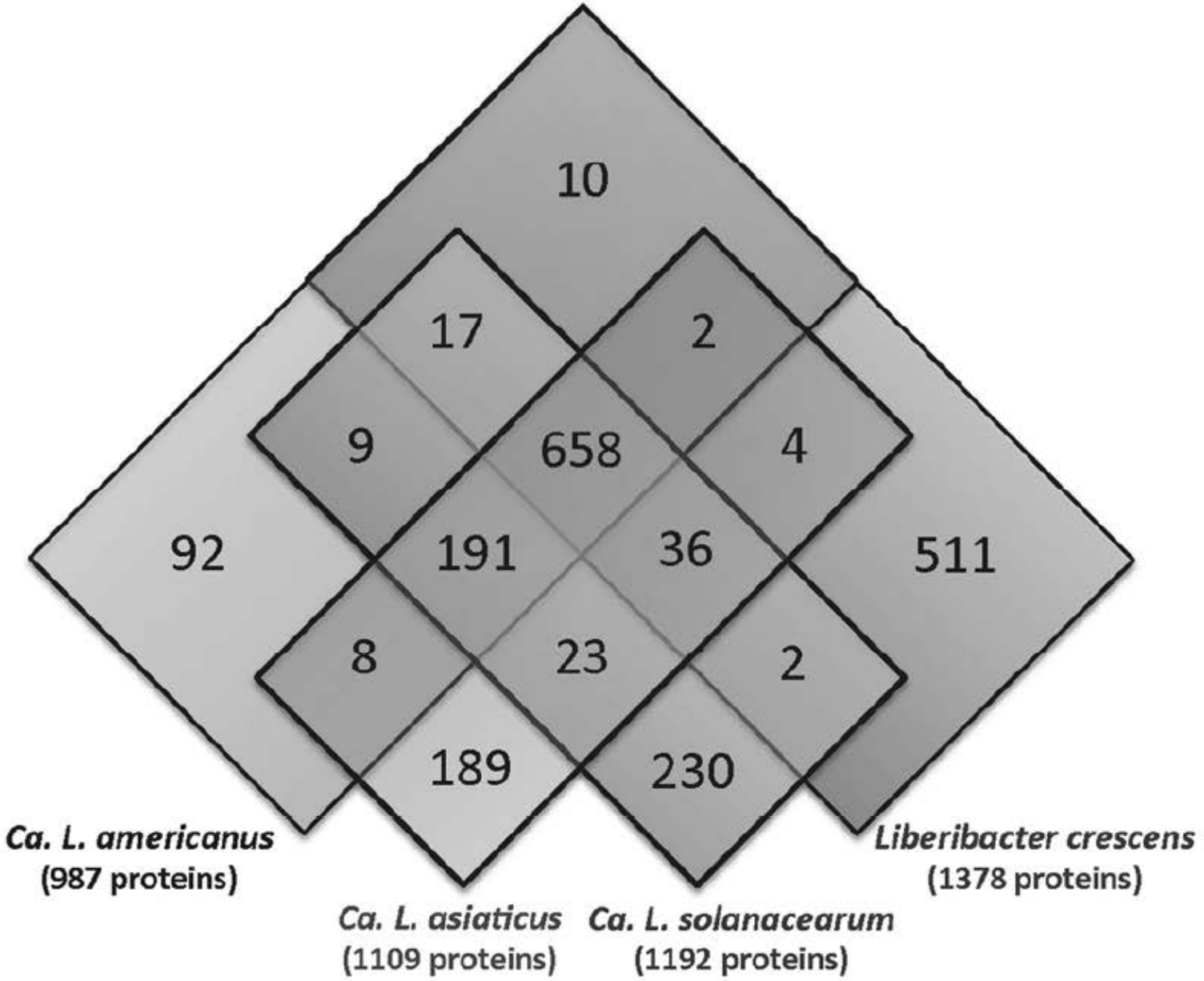


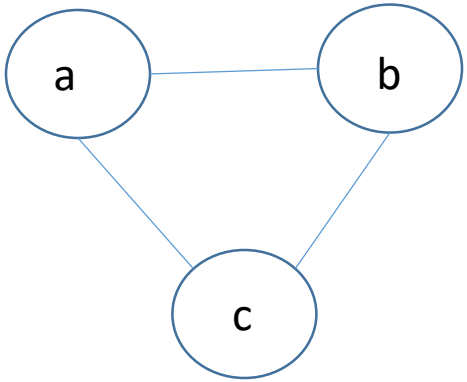
Fig. 2. Venn diagram showing protein-coding gene sharing among the four Liberibacter genomes. Numbers below each species are the total number of predicted protein-coding genes for that genome.

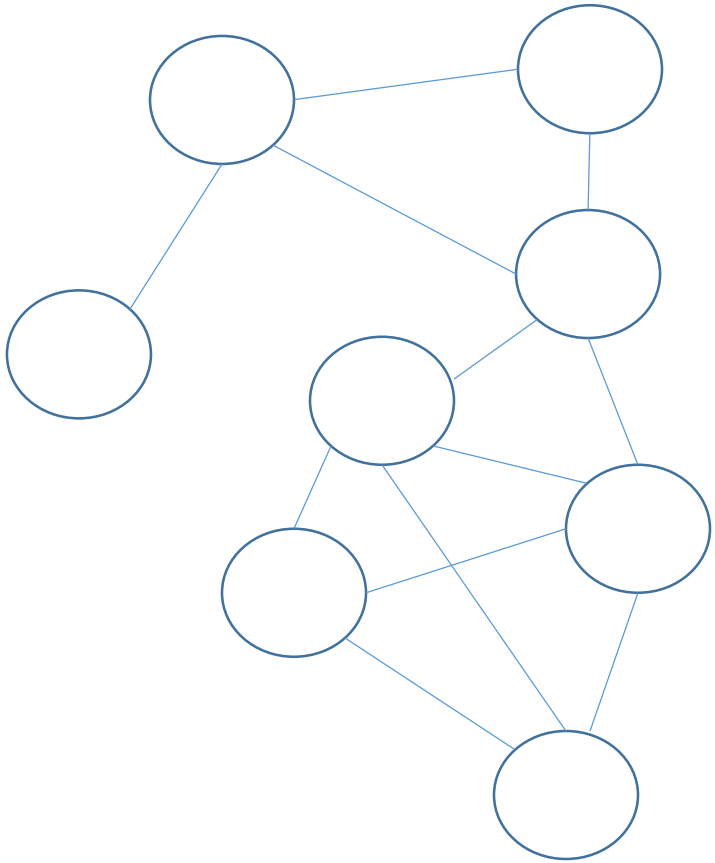
Wulff et al. MPMI Vol. 27, No. 2, 2014, pp. 163–176 http://dx.doi.org/10.1094/MPMI-09-13-0292-R.

Cômputo de famílias de proteínas

1. Verificar as similaridades entre as sequências
   a) Usando (por exemplo) BLAST + critérios
2. Representar as similaridades num grafo
3. Aplicar um algoritmo de clusterização sobre o grafo

Clusterização é necessária porque o grafo pode ser complexo





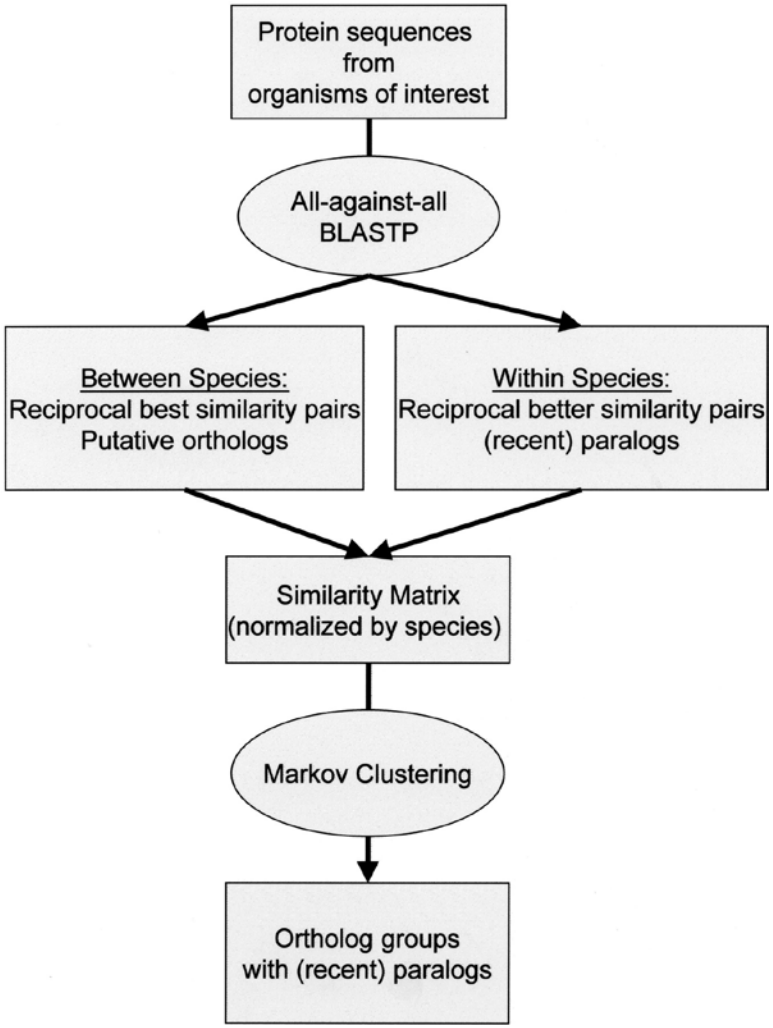
Resultado da clusterização

- Matriz com genes nas colunas (g_j), genomas nas linhas (O_i)
- Cada coluna representa uma família de genes homólogos

	g1	g2	g3	g4	…
O1	✓		✓	✓	
O2		✓	✓		
O3	✓		✓		
O4			✓	✓	✓
O5	✓		✓		✓
O6			✓		
…			✓		

Foi este tipo de processamento que permitiu obter dados para construir a árvore da vida vista na primeira aula do tema “comparação de sequências”

orthoMCL pipeline



Li Li et al. Genome Res. 2003; 13: 2178-2189

NEWS

Mar 31, 2011
OrthoMCL-DB Version 5 is released. We have included 150 genomes in this release.

May 31, 2010
OrthoMCL-DB Version 4 is released. There are 138 genomes included in version 4.

Oct 9, 2009
OrthoMCL-DB Version 3 is released. The new dataset includes 128 genomes. New web site features include: (1) a tool to assign your proteins to OrthoMCL groups (see the new tools menu); (2) a mapping from Version 3 groups to Version 2 and 1 is available for searching, allowing you to track changes across versions; (3) the phyletic pattern in the groups result page is configurable, so you can tailor it to the clades you are interested in.

Sep 22, 2009
OrthoMCL-DB pipeline re-engineered. We have completely overhauled how we produce the database, encoding the entire process in a pipeline system. This should dramatically improve our ability to deliver new versions of the database.

Feb 28, 2008

Welcome to OrthoMCL DB

Ortholog Groups of Protein Sequences from Multiple Genomes!

Current Release:
Version: 5
Number of Genomes: 150
Number of Protein Sequences: 1398546
Number of Ortholog Groups: 124740

Search for Groups

- by IDs, Keyword, or PFam domain
- by Phyletic Pattern
- by Phyletic Pattern - Advanced
- by Group Properties
- Query History - Groups

Search for Sequences

- by IDs, Keyword, Taxonomy or PFam domain
- by BLAST Search
- Query History - Sequences

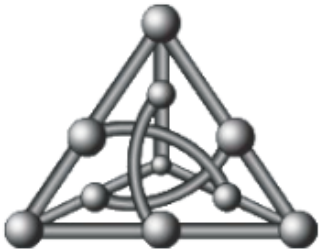
Tools

- Assign your proteins to OrthoMCL groups

OrthoMCL Software

Database Download

Xanthomonas fuscans str. *aurantifolii* Genome Project



Choose genomes to see OrthoMCL families containing genes from them

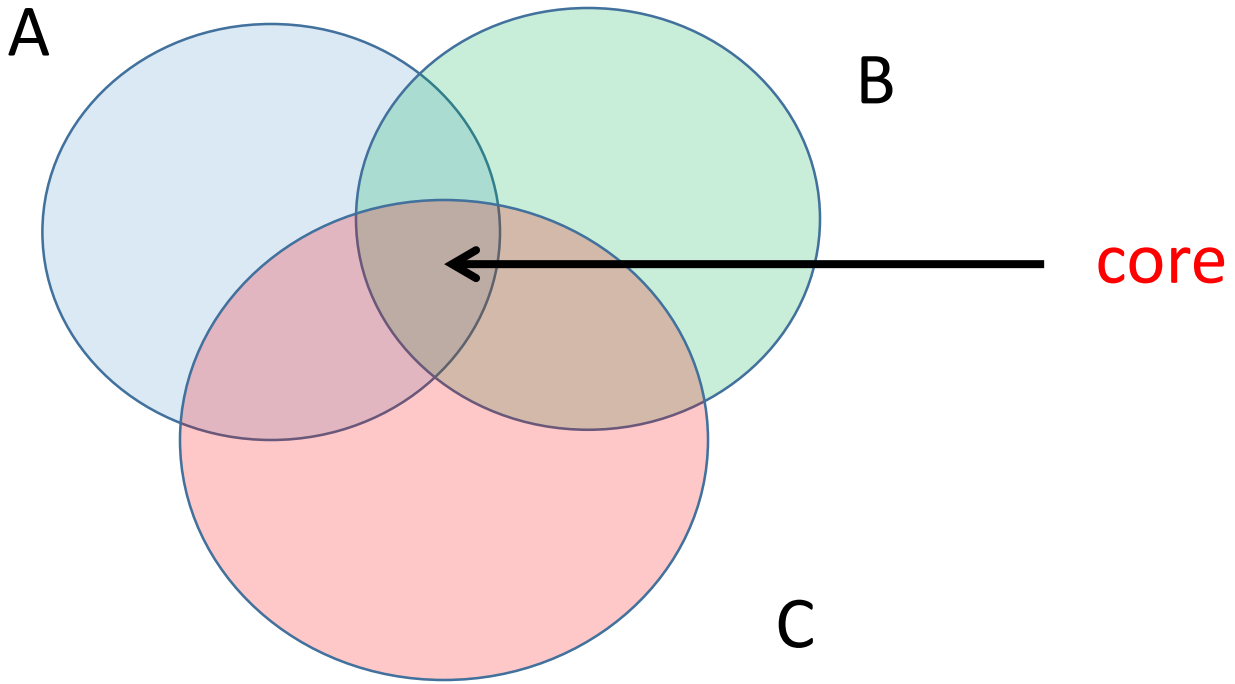
Include	Exclude	I don't care	
○	○	◉	Xanthomonas fuscans subsp. aurantifolii str. 11122 (B)
○	○	◉	Xanthomonas fuscans subsp. aurantifolii str. 10535 (C)
○	○	◉	Xanthomonas campestris pv. campestris str. 8004
○	○	◉	Xanthomonas campestris pv. campestris str. ATCC 3391
○	○	◉	Xanthomonas campestris pv. campestris B100
○	○	◉	Xanthomonas campestris pv. vesicatoria str. 85-10
○	○	◉	Xanthomonas axonopodis pv. citri str. 306
○	○	◉	Xanthomonas oryzae pv. oryzae KACC10331
○	○	◉	Xanthomonas oryzae pv. oryzae MAFF 311018
○	○	◉	Xanthomonas oryzae pv. oryzae PXO99A
○	○	◉	Xanthomonas albilineans GPE PC73
○	○	◉	Xylella fastidiosa 9a5c
○	○	◉	Xylella fastidiosa M12
○	○	◉	Xylella fastidiosa M23 plasmid pXFAS01
○	○	◉	Xylella fastidiosa Temecula1

○ check all ○ check all ○ check all(see all families)

Submit Query Reset

```
========================================
>Family 5286  -- Gblocks 331/384 (86%)
21243924              nr   alkanesulfonate transporter substrate-binding subunit [Xanthomonas axonopodis pv. citri str. 306]
9810940               nr   - nitrate transport protein [Xanthomonas fuscans subsp. aurantifolii str. 11122 (B)]
9929350               nr   - nitrate transport protein [Xanthomonas fuscans subsp. aurantifolii str. 10535 (C)]

count_ = 3
========================================
>Family 5287  -- Gblocks 346/346 (100%)
21243925              nr   oxidoreductase [Xanthomonas axonopodis pv. citri str. 306]
9810930               nr   - oxidoreductase [Xanthomonas fuscans subsp. aurantifolii str. 11122 (B)]
9929360               nr   - oxidoreductase [Xanthomonas fuscans subsp. aurantifolii str. 10535 (C)]

count_ = 3
========================================
>Family 5288  -- Gblocks 442/442 (100%)
21243926              nr   nitrilotriacetate monooxygenase component A [Xanthomonas axonopodis pv. citri str. 306]
9810920               nr   - nitrilotriacetate monooxygenase component A [Xanthomonas fuscans subsp. aurantifolii str. 11122 (B)]
9929370               nr   - nitrilotriacetate monooxygenase component A [Xanthomonas fuscans subsp. aurantifolii str. 10535 (C)]

count_ = 3
========================================
>Family 5289  -- Gblocks 356/356 (100%)
21243950              nr   avirulence protein [Xanthomonas axonopodis pv. citri str. 306]
9814680               nr   - type III secretion system hopXl-like protein [Xanthomonas fuscans subsp. aurantifolii str. 11122 (B)]
9900040               nr   - type III secretion system hopXl-like protein [Xanthomonas fuscans subsp. aurantifolii str. 10535 (C)]

count_ = 3
========================================
>Family 5290  -- Gblocks 291/297 (97%)
21243956              nr   hypothetical protein XAC3230 [Xanthomonas axonopodis pv. citri str. 306]
9826830               nr   - conserved hypothetical protein [Xanthomonas fuscans subsp. aurantifolii str. 11122 (B)]
9923780               nr   - conserved hypothetical protein [Xanthomonas fuscans subsp. aurantifolii str. 10535 (C)]

count_ = 3
========================================
>Family 5291  -- Gblocks 203/373 (54%)
21243959              nr   transposase [Xanthomonas axonopodis pv. citri str. 306]
9803570               nr   - transposase [Xanthomonas fuscans subsp. aurantifolii str. 11122 (B)]
9900450               nr   - transposase [Xanthomonas fuscans subsp. aurantifolii str. 10535 (C)]

count_ = 3
========================================
>Family 5292  -- Gblocks 163/171 (95%)
21243960              nr   hypothetical protein XAC3234 [Xanthomonas axonopodis pv. citri str. 306]
9820110               nr   - conserved hypothetical protein [Xanthomonas fuscans subsp. aurantifolii str. 11122 (B)]
```

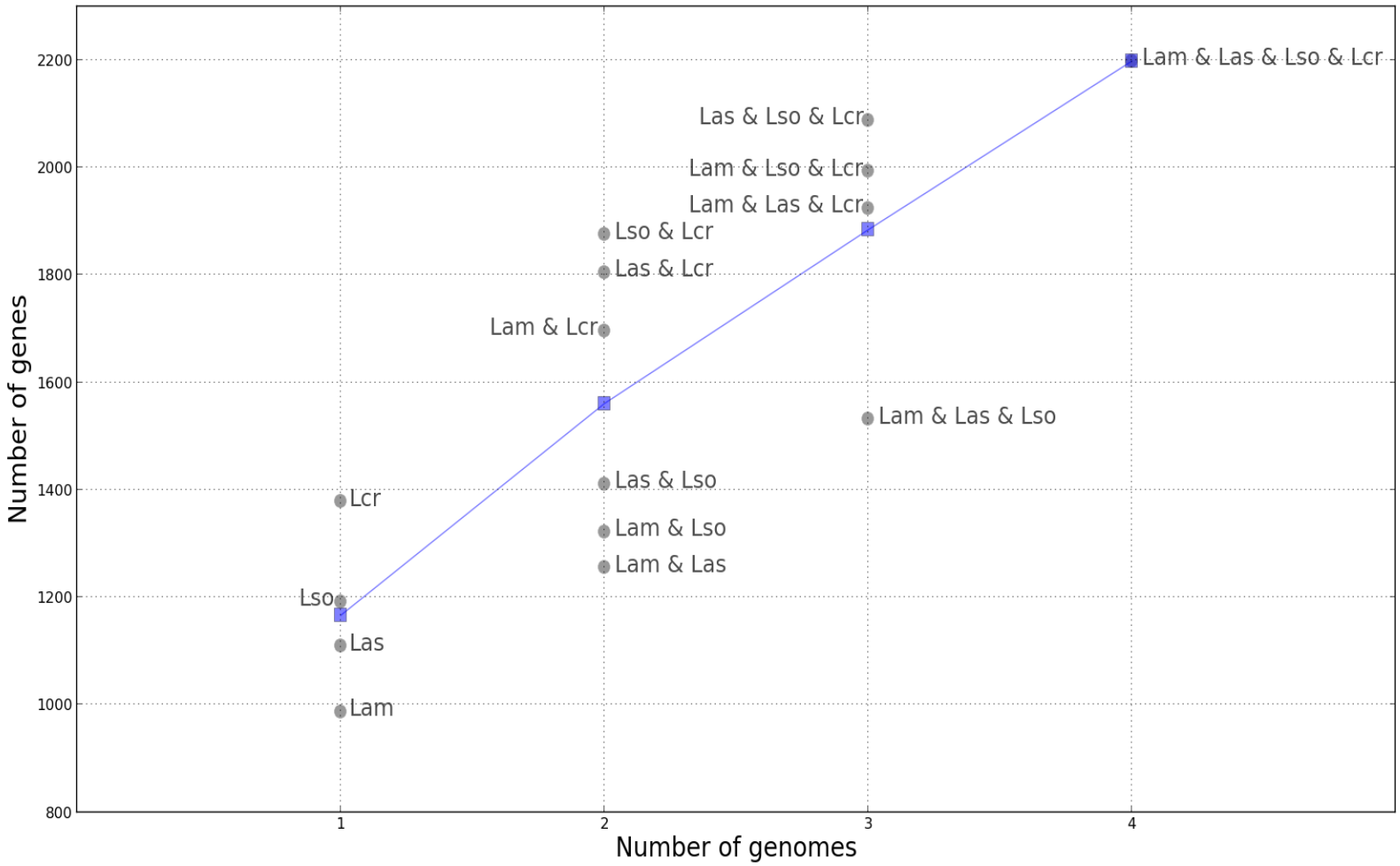
Pan genoma; genoma core e genoma acessório



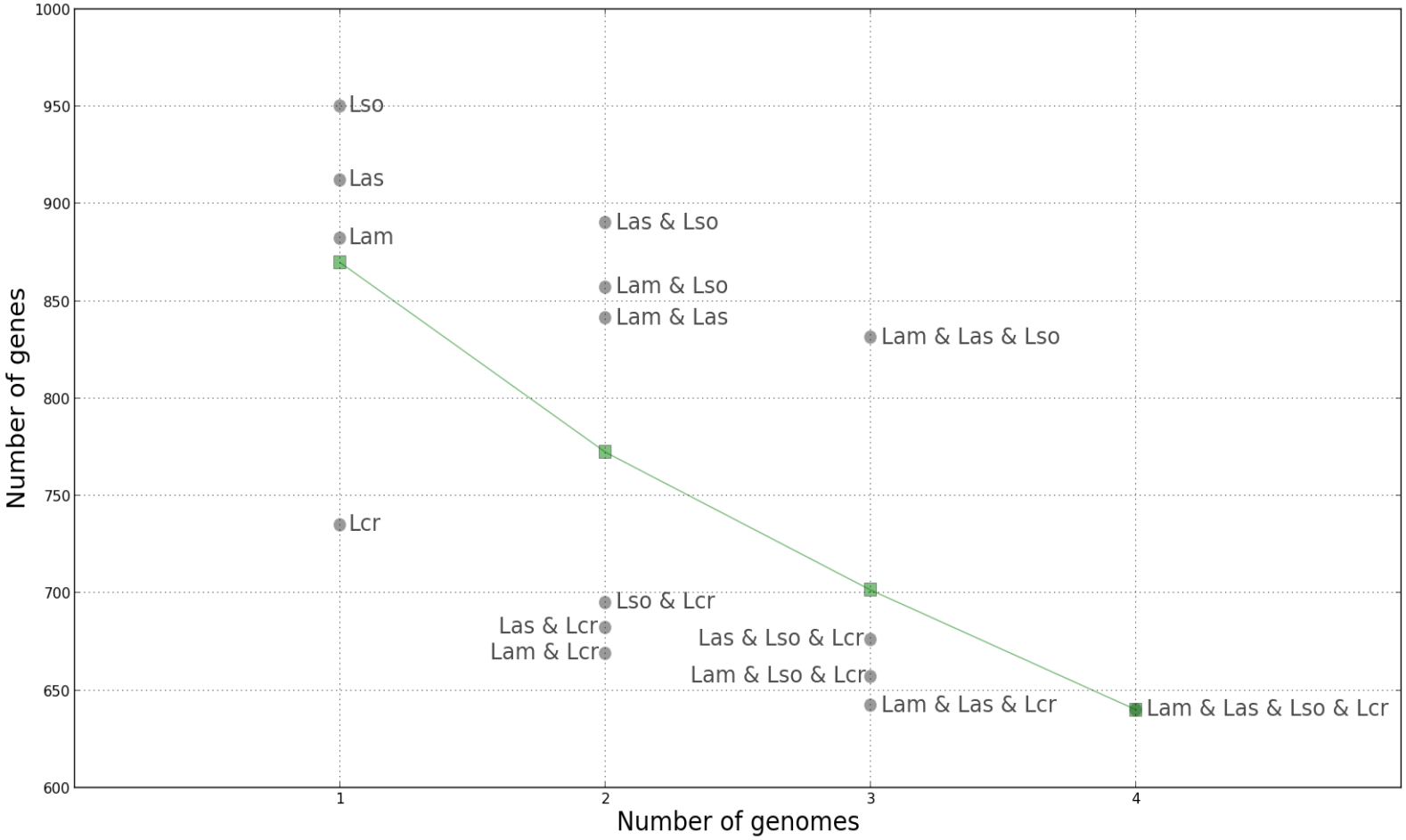
pan: A U B U C

acessório: pan menos core

Curva de pan-genoma (n = 4)



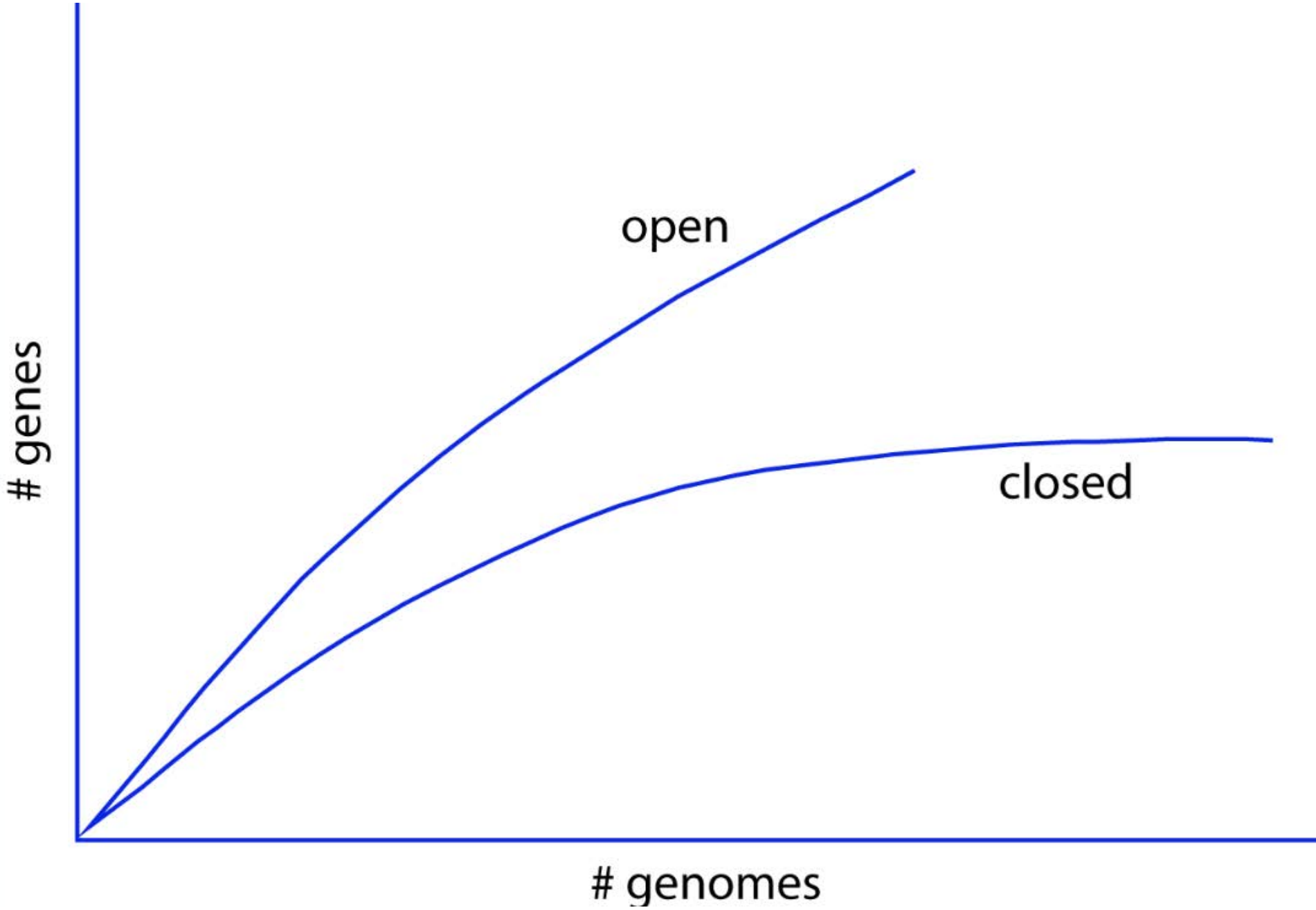
Curva de core genoma (n = 4)



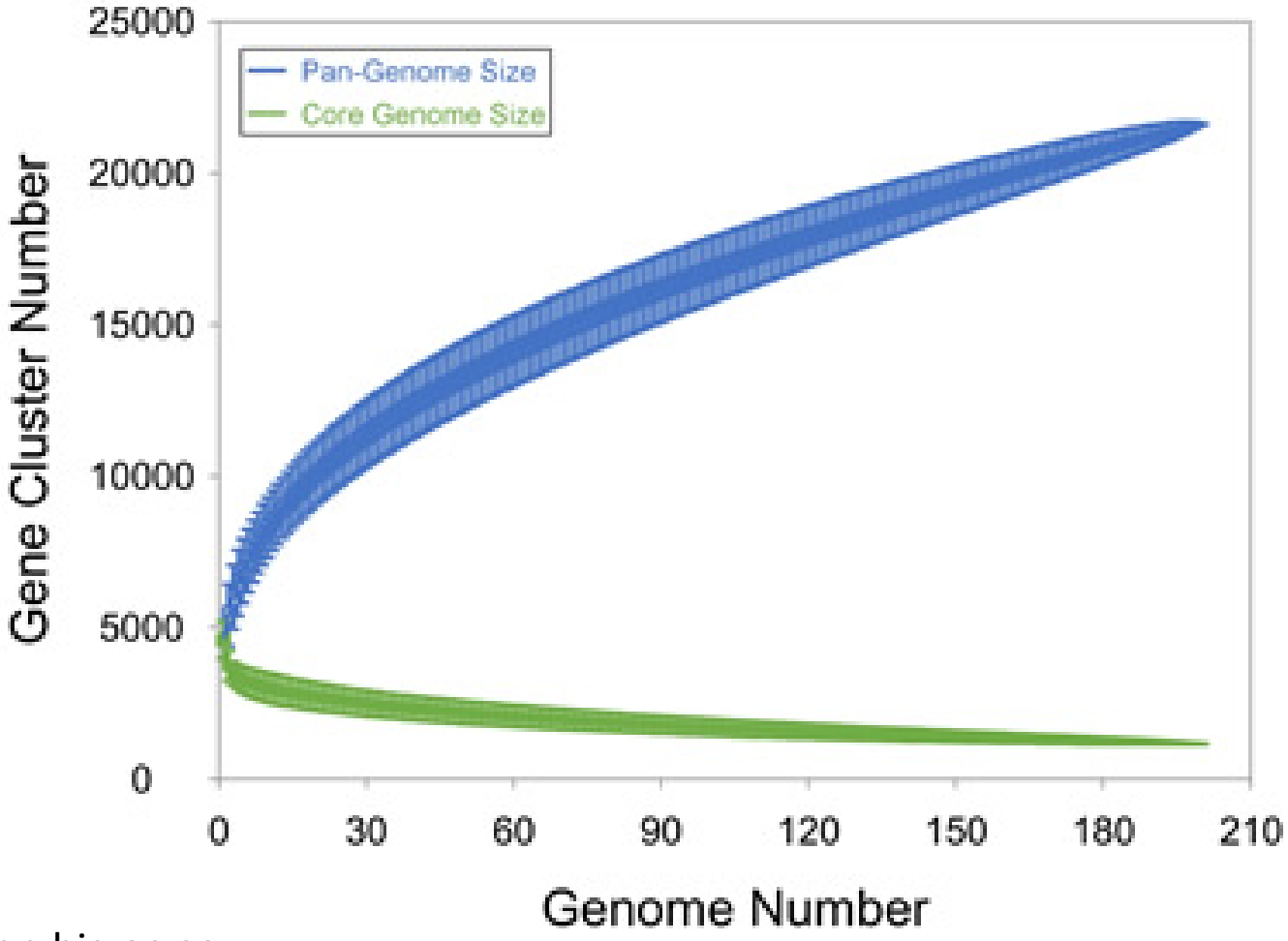
O número de genes para x=1 não é o mesmo do gráfico anterior pois os singletons são descartados

Genomas fechados e abertos

- Fechado
  - O número de genes/famílias do pan-genoma atinge um platô
- Aberto
  - O número de genes/famílias não atinge um platô; cresce sempre que se acrescentam novos genomas


open
closed
genes
genomes

O genoma de *E. coli* é aberto



Fonte: pangp.big.ac.cn

Bacillus anthracis

- Genoma fechado
- Cerca de 3.000 genes

Ferramentas para análise pan/core

- Get_Homologues
- Contreras-Moreira, B. and P. Vinuesa, *GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis.* Appl Environ Microbiol, 2013. 79(24): p. 7696-701

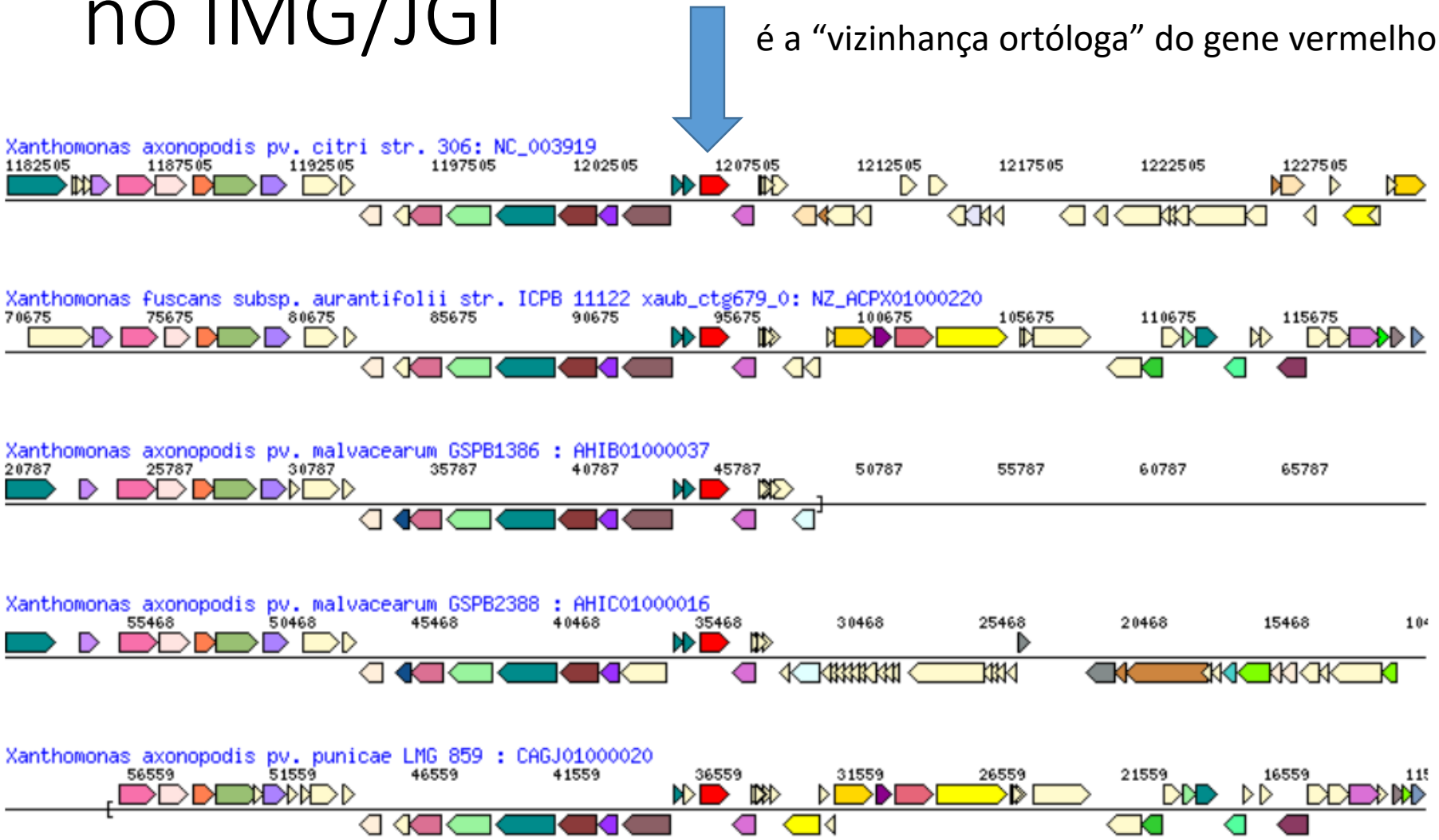
- Panaroo
- Tonkin-Hill, G., MacAlasdair, N., Ruis, C. et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. Genome Biol 21, 180 (2020). https://doi.org/10.1186/s13059-020-02090-4

Conjuntos + contexto

- Como genes compartilhados aparecem em seus respectivos genomas?
- **Filogenômica**
- Busca de sintenia = preservação de ordem
- Basta fazer um alinhamento
  - Os “caracteres” a serem alinhados são os genes
  - *alinhamento de ortólogos*

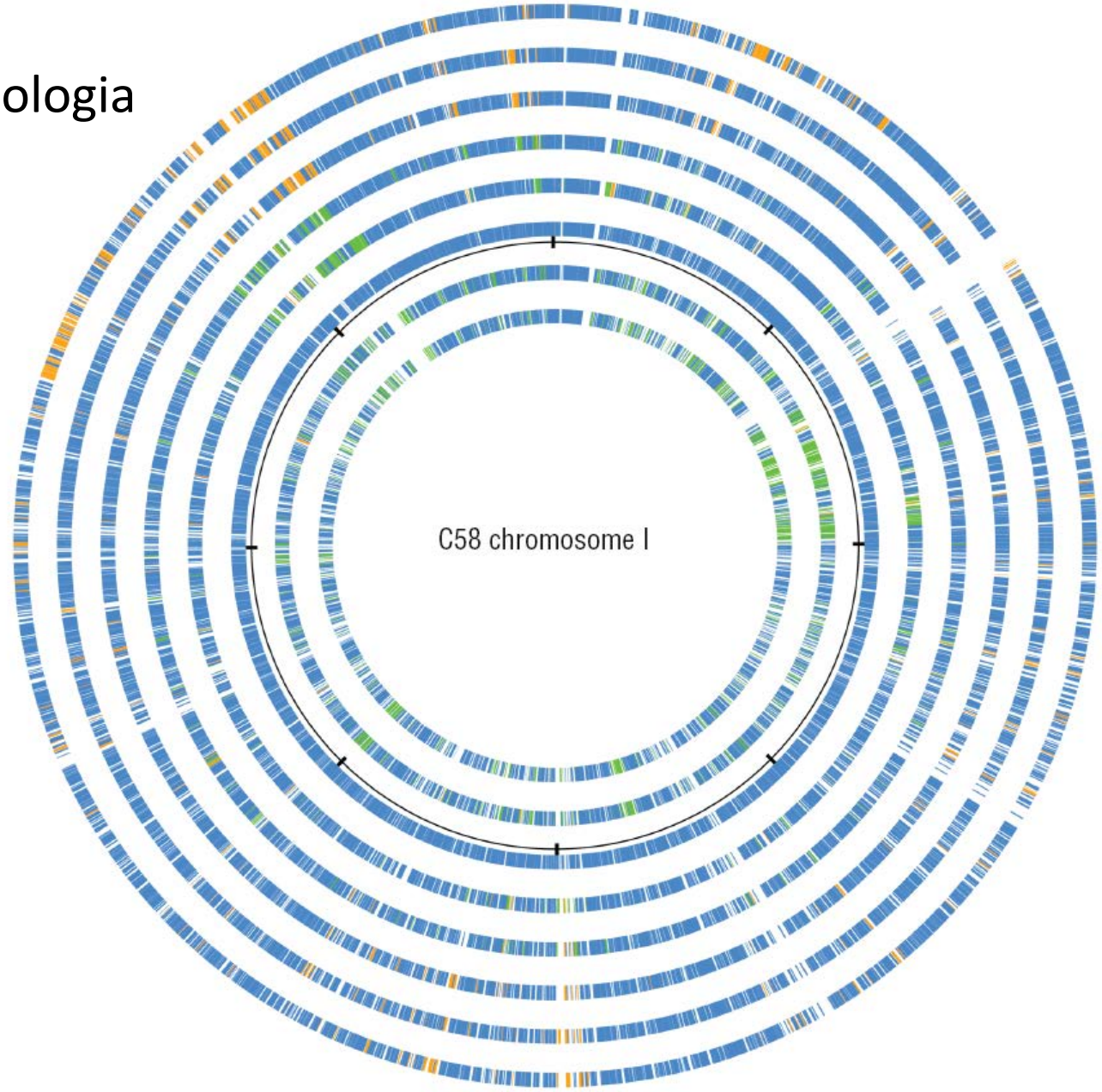
Alinhamento de ortólogos obtido no IMG/JGI

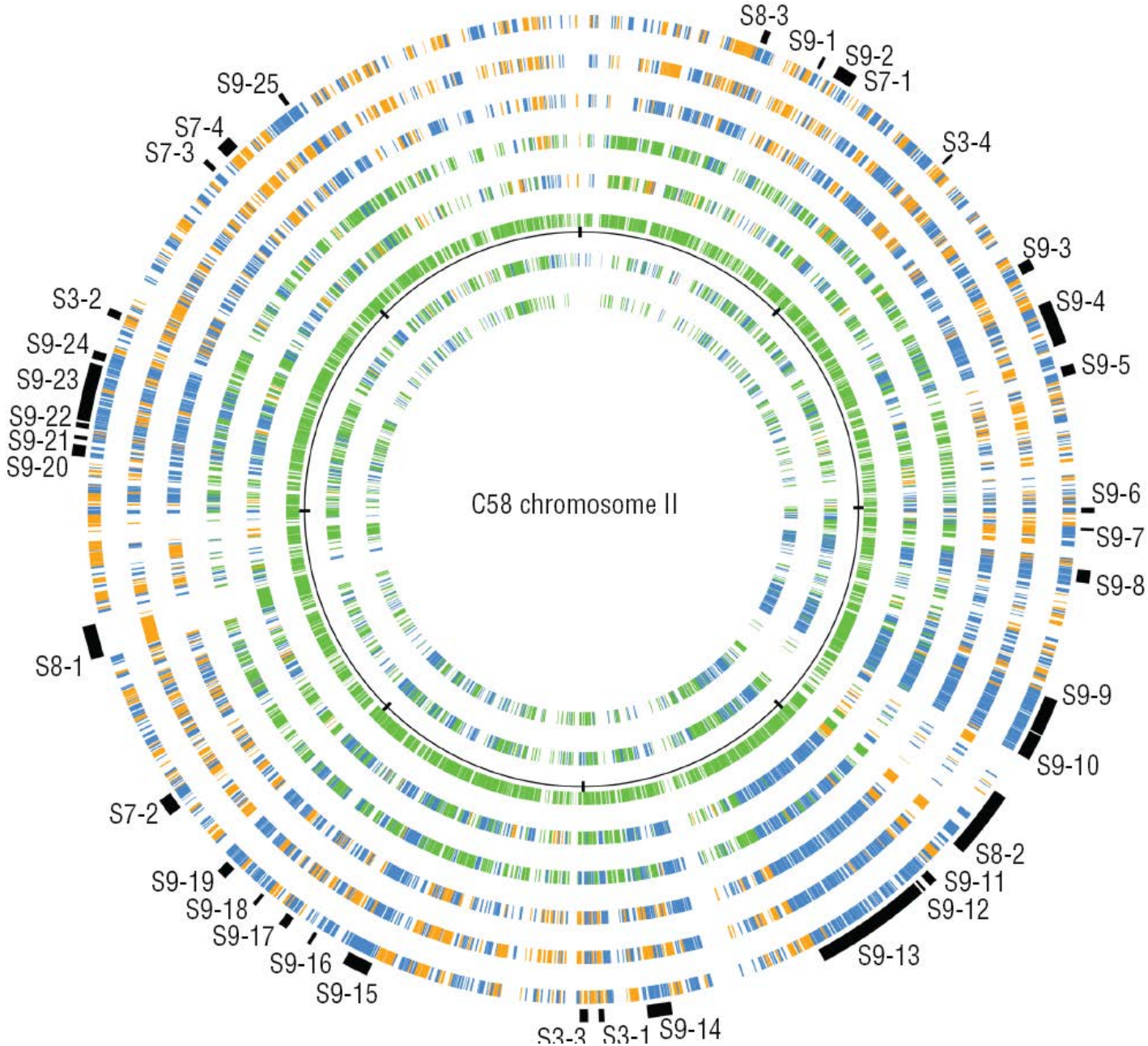
é a “vizinhança ortóloga” do gene vermelho



genes de mesmas cores são ortólogos

Roda da ortologia




C58 chromosome II
S8-3
S9-1
S9-2
S7-1
S3-4
S9-3
S9-4
S9-5
S9-6
S9-7
S9-8
S9-9
S9-10
S8-2
S9-11
S9-12
S9-13
S9-14
S3-1
S3-3
S9-15
S9-16
S9-17
S9-18
S9-19
S7-2
S8-1
S9-20
S9-21
S9-22
S9-23
S9-24
S3-2
S9-25
S7-3
S7-4

Protein family resources

Clusters of orthologous groups (COG, KOG, eggNOG)

KEGG orthologs

wellcome trust sanger institute

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam 26.0 (November 2011, 13672 families)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. **More...**

Pfam

keyword search Go

QUICK LINKS	YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...
SEQUENCE SEARCH	Analyze your protein sequence for Pfam matches
VIEW A PFAM FAMILY	View Pfam family annotation and alignments
VIEW A CLAN	See groups of related families
VIEW A SEQUENCE	Look at the domain organisation of a protein sequence
VIEW A STRUCTURE	Find the domains on a PDB structure
KEYWORD SEARCH	Query Pfam by keywords
JUMP TO	enter any accession or ID Go Example

Enter any type of accession or ID to jump to the page for a Pfam family or clan, UniProt sequence, PDB structure, etc.

Or view the help pages for more information

Recent Pfam blog posts

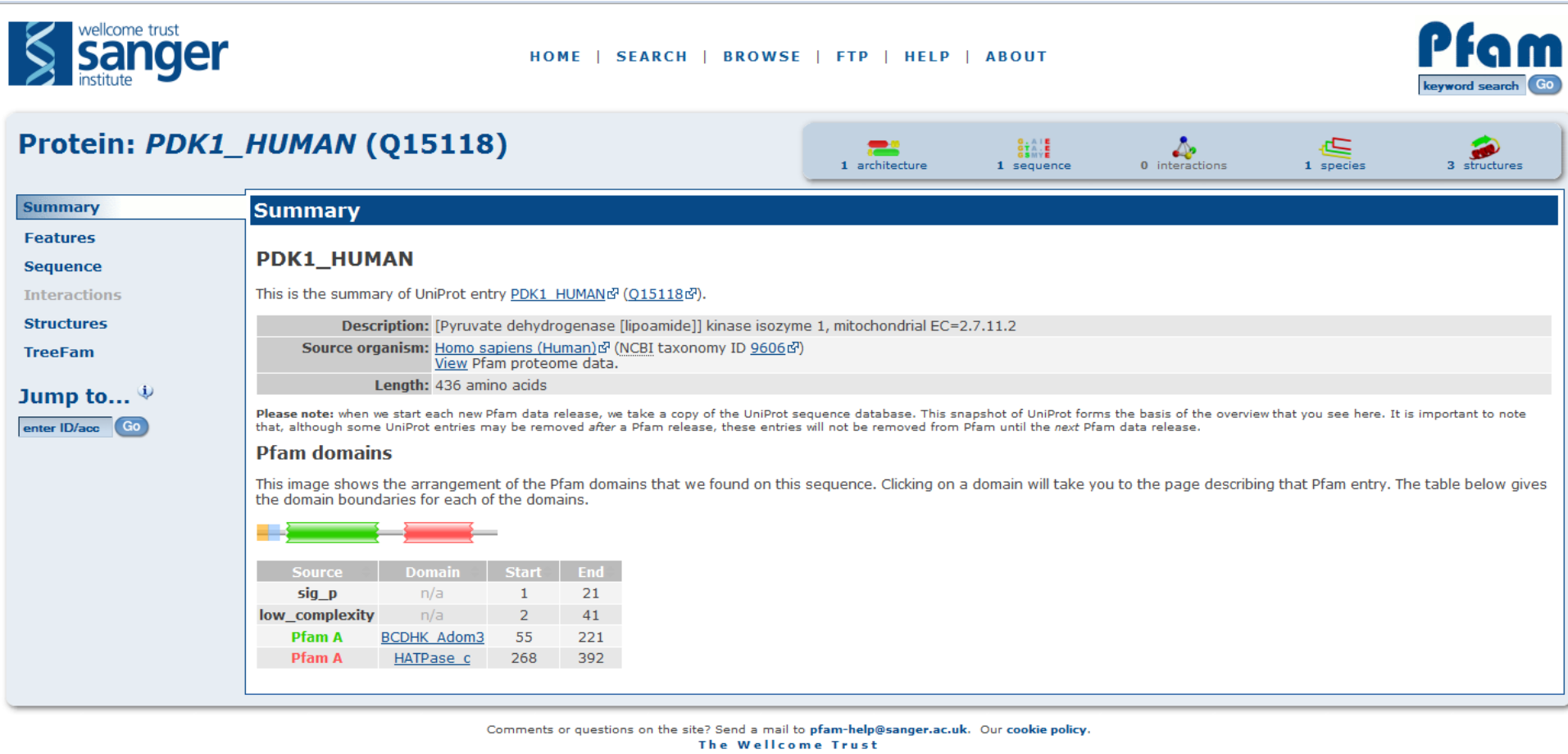
☒Hide th

Does my family of interest have a determined 3D protein structure? (posted 9 May 2012)

Two related questions that we are often asked via the Pfam helpdesk is 'Which families have a known three-dimensional structure?' and 'Why is a particular a PDB structure not found in Pfam'. You may thir that there are obvious answers to these questions – but as with many things in life the answer is not [...]

TreeFam is back with a new release ! (posted 27 March 2012)

Query by accession



Quick links

- Whole genome function views
- Gene expression tools
- cSNP tools
- Upload multiple gene IDs
- Community Curation
- My Workspace
- HMM scoring
- Downloads
- Genome statistics
- Site map

Newsletter subscription

Enter your Email:

Subscribe

KEYWORD SEARCH

All Go

SEQUENCE SEARCH

Enter a protein sequence:

Sequence query limits: Protein - 50kb

Submit

The PANTHER (**P**rotein **AN**alysis **TH**rough **E**volutionary **R**elationships) Classification System is a unique resource that **classifies genes by their functions**, using published scientific experimental evidence and evolutionary relationships to predict function even in the absence of direct experimental evidence. Proteins are **classified by expert biologists** according to:

- Gene families and subfamilies, including annotated phylogenetic trees
- Gene Ontology classes: molecular function, biological process, cellular component
- PANTHER Protein Classes
- Pathways, including diagrams

PANTHER is part of the Gene Ontology Reference Genome Project.

PANTHER is supported by a research grant from the National Institute of General Medical Sciences [grant GM081084] and maintained by the Thomas lab at the University of Southern California.

What can I do on the PANTHER site?
Guide to getting started

News

(March 16, 2012)

PANTHER 7.2 is released.

Click for additional info.

Publications

How to cite PANTHER

"PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium." Mi, et al.

"Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools." Thomas, et al.

"PANTHER: a library of protein families and subfamilies indexed by function." Thomas, et al.

PANTHER HMM SEQUENCE SCORING RESULTS

The top scoring HMM is reported, along with the E-value (the number of expected false-positive hits expected). If the E-value is less than 1e-3, no hits are reported.

PANTHER Hit: PROLINE SYNTHETASE CO-TRANSCRIBED BACTERIAL HOMOLOG PROTEIN (PTHR10146)
HMM E-value score: 2.2e-113 ●●●

```
Sequence Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
-------- ------- ----- -----    ----- -----      -----  -------
sequence   1/1       1   234 []    11   253 ..   387.5 2.2e-113

Alignments of top-scoring domains:
sequence: domain 1 of 1, from 1 to 234: score 387.5, E = 2.2e-113
                     *->lgvaanLakVlerikaaaakagRdppavrLvAVSKTkPaelileayd
                        ++va+nL++V++++++a+ak+ R   + +LvAVSKTkP+e+++eay+
    sequence     1     MAVAKNLLAVRAKVAEAVAKSARQ-QQCTLVAVSKTKPVEDLQEAYE 46

                     aGqRhFGENYvQElleKaplLpdlcpdikWHFIGhLQsNKvkkll.gvpn
                     a qRhFGENY+QEl++KaplLp    d+kWH+IGh+QsNK+k+l+++vpn
    sequence    47 ADQRHFGENYIQELVQKAPLLPK---DVKWHYIGHVQSNKAKPLVrDVPN 93

                     ldmvhsvDslklAdklnkaaaklkglgkplkvlvQVNtsGEesKsGvppe
                     l++v++vDs+k+A++lnka  ++  ++++l+v+vQVNts Ee+KsG++ +
    sequence    94 LFVVETVDSIKIANALNKASGEF--RSEKLNVMVQVNTSEEEQKSGIDAD 141

                     Elpellkhvlkkcpn LellGLMTIGpfdgdlekgpnpdFalLaklrkevc
                       +el++h+++ c++L+l GLMTIG++++  ++     F +L+++rk+v+
    sequence   142 GSVELAQHIVSSCEHLNLTGLMTIGRYGDTTSE----CFDRLVACRKRVA 187

                     kklglnpkllELSMGMSgDfelAIeaGsTlVRvGsaIFGeRdypkkp<-*
                     +++g  +  l LSMGMSgDfelAI  GsT+VRvGs+IFG+R+y +k+
    sequence   188 EAIGKAETDLDLSMGMSGDFELAISCGSTHVRVGSTIFGARNYANKE    234
```

Search

PANTHER families

Go

Quick links

- Whole genome function views
- Gene expression tools
- cSNP tools
- Upload multiple gene IDs
- Community Curation
- My Workspace
- HMM scoring
- Downloads
- Genome statistics
- Site map

PANTHER FAMILY INFORMATION

Family:	PROLINE SYNTHETASE CO-TRANSCRIBED BACTERIAL HOMOLOG PROTEIN (PTHR10146)
Subfamilies:	1
PANTHER Links:	Tree MSA
GO Molecular Function:	catalytic activity
GO Biological Process:	metabolic process ↳ primary metabolic process ↳ cellular amino acid and derivative metabolic process ↳ cellular amino acid metabolic process
GO Cellular Component:	
PANTHER protein class:	
Pathway Categories:	No pathway information available
Genes:	45
HMM Length	273
Downloads:	HMM (HMMER format)

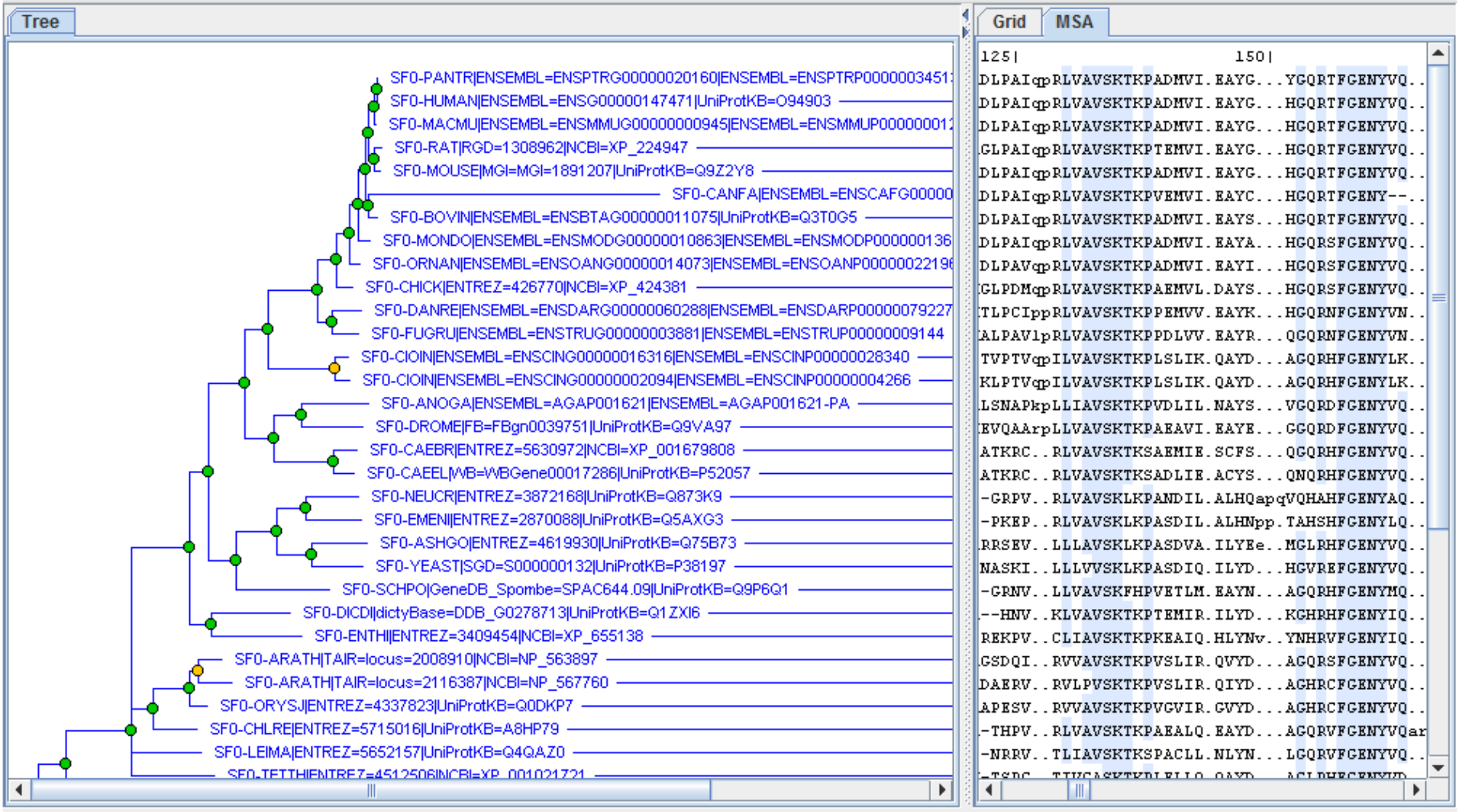
GENES ASSIGNED TO THIS FAMILY

Species	Count
Anopheles gambiae	1
Aquifex aeolicus vf5	1
Arabidopsis thaliana	2

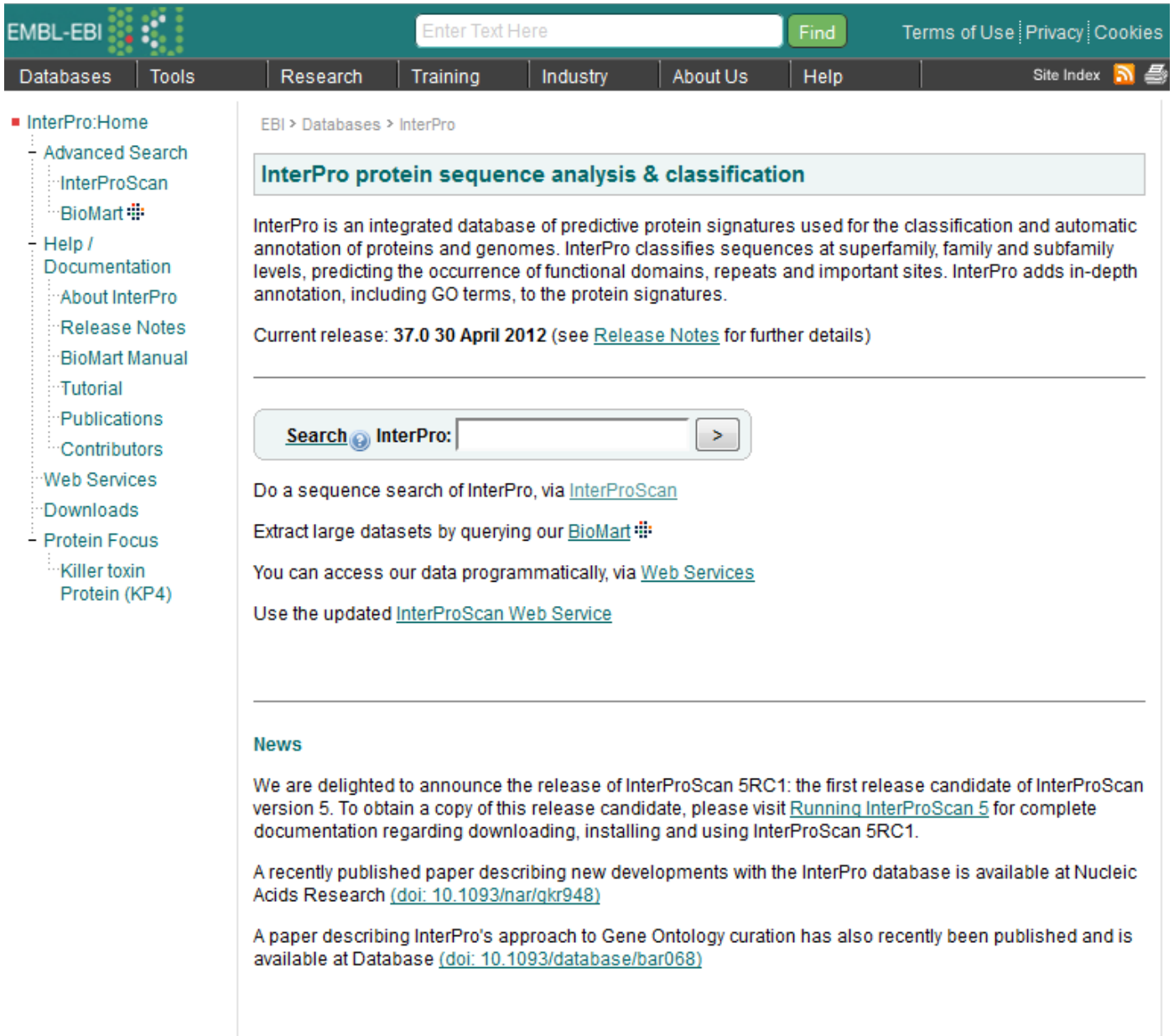
PANTHER TREE VIEWER • close window

Family Name : PROLINE SYNTHETASE CO-TRANSCRIBED BACTERIAL HOMOLOG PROTEIN (PTHR10146)

Tree MSA



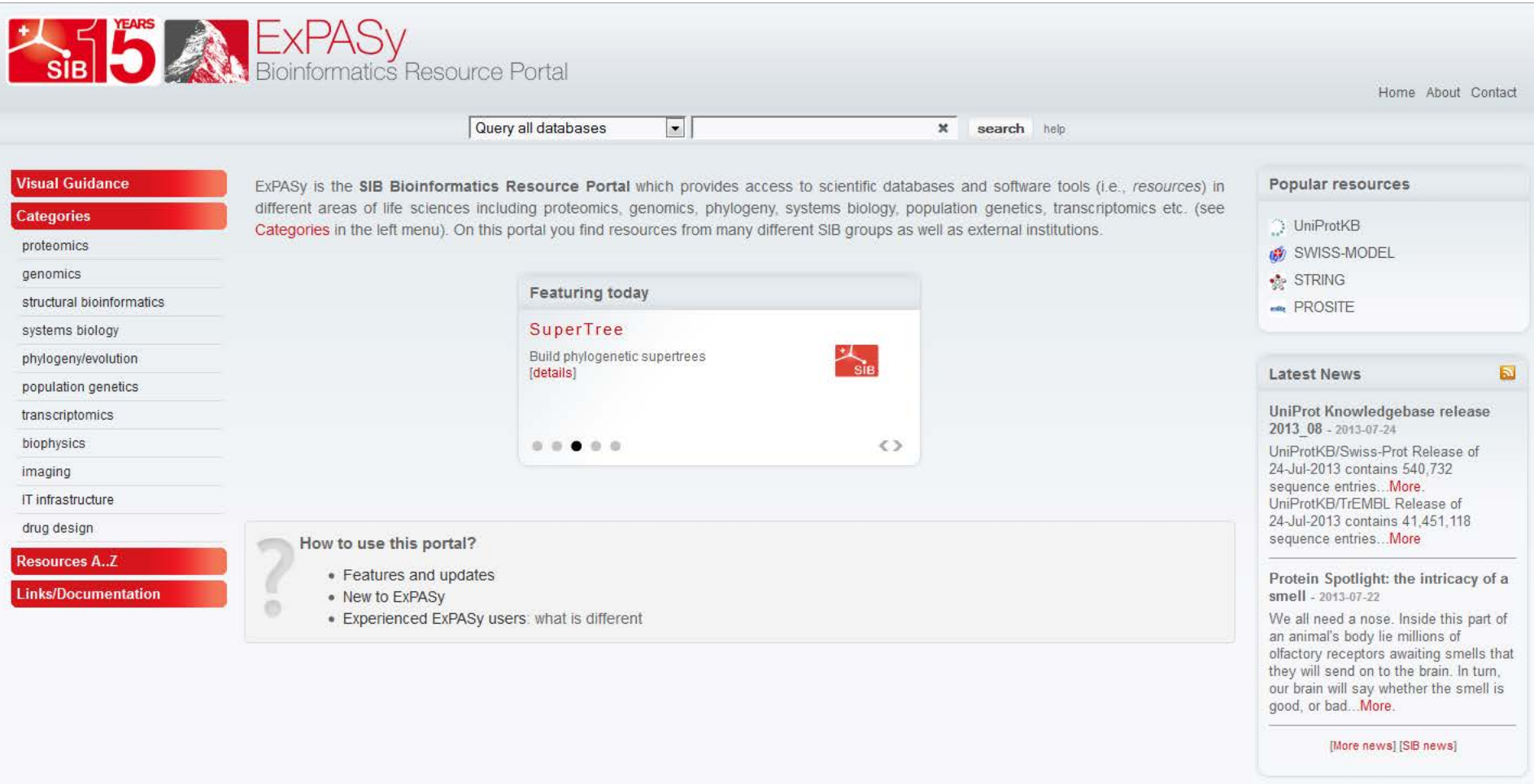
Resource federation: InterPro



Not as easy as it may sound…

- Specific protein families may not be consistent across resources
- Most families (MSAs, trees, HMMs) in these resources are not manually curated
  - Domains in Pfam-A are curated
  - TIGRfams are curated
  - HAMAP families are curated

http://www.expasy.org



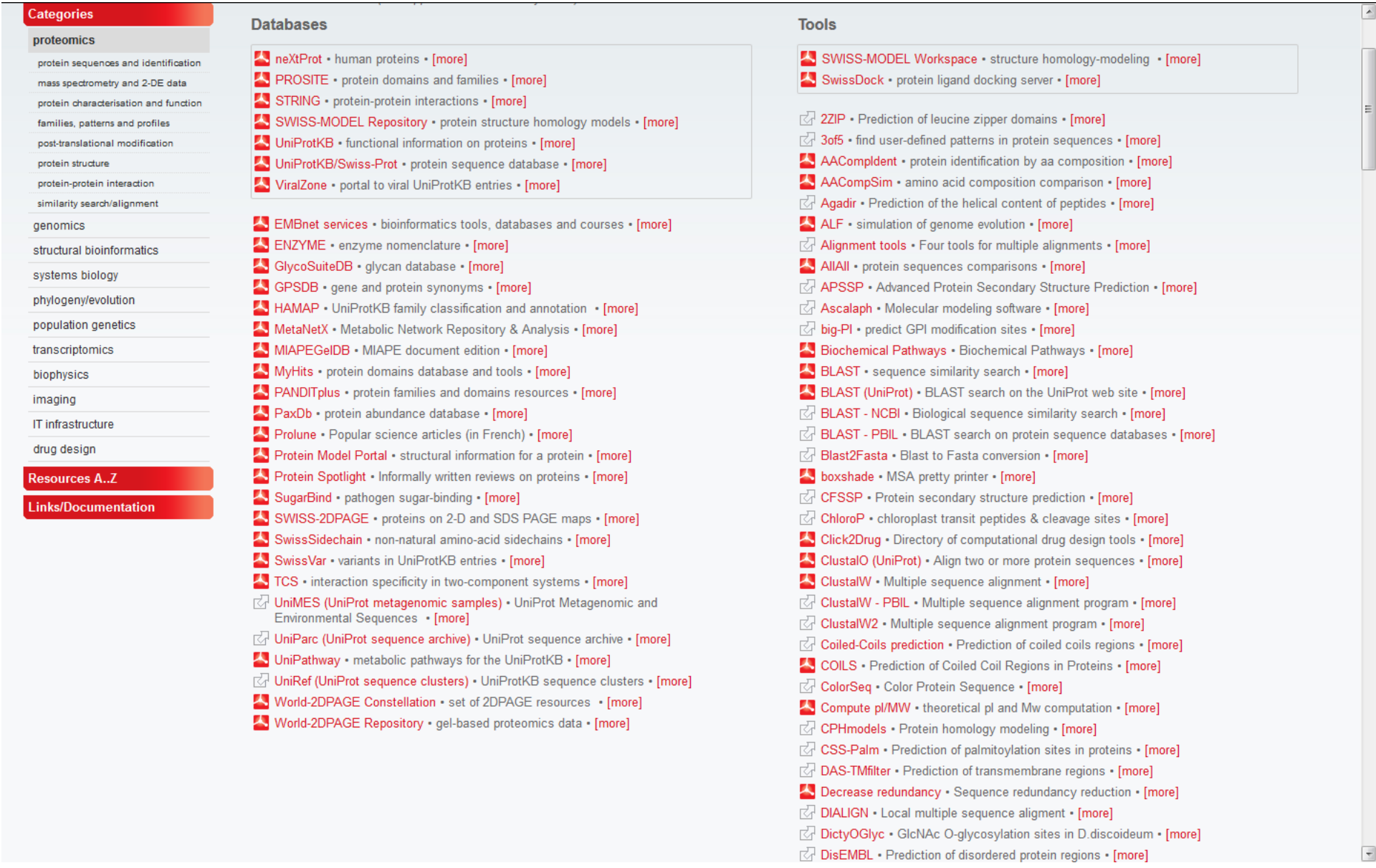
Visual Guidance
Categories
proteomics
genomics
sequence alignment
similarity search
characterisation/annotation
structural bioinformatics
systems biology
phylogeny/evolution
population genetics
transcriptomics
biophysics
imaging
IT infrastructure
drug design
Resources A..Z
Links/Documentation

SIB resources
External resources - *(No support from the ExPASy Team)*

Tools

- Alignment tools • Four tools for multiple alignments • [more]
- boxshade • MSA pretty printer • [more]
- ClustalW • Multiple sequence alignment • [more]
- ClustalW - PBIL • Multiple sequence alignment program • [more]
- ClustalW2 • Multiple sequence alignment program • [more]
- Codon Suite • codon-based sequence analysis • [more]
- Decrease redundancy • Sequence redundancy reduction • [more]
- DIALIGN • Local multiple sequence aligment • [more]
- GENIO/logo • RNA/DNA & Amino Acid Sequence Logos • [more]
- Kalign - EBI • Fast and accurate multiple sequence alignment • [more]
- Kalign - SBC • Fast and accurate multiple sequence alignment • [more]
- LALIGN • Pairwise alignment • [more]
- MADAP • clustering for genome annotation data • [more]
- MAFFT - CBRC • Multiple sequence aligment • [more]
- MAFFT - EBI • Multiple sequence aligment • [more]
- MaxAlign • Gap removal from alignments • [more]
- Multialin • Multiple sequence alignment • [more]
- MUSCLE • Multiple alignment server • [more]
- Newick Utilities • high-throughput phylogenetic tree processing • [more]
- Phylogibbs • regulatory sites discovery • [more]
- SIBsim4 • spliced sequence alignment • [more]
- T-Coffee • sequence and structure multiple alignments • [more]
- T-Coffee - EBI • Multiple sequence alignment program • [more]
- T-Coffee - WUR • Multiple sequence alignment program • [more]
- WebLogo • Sequence logos • [more]

proteômica



Categories

- proteomics
  - protein sequences and identification
  - mass spectrometry and 2-DE data
  - protein characterisation and function
  - families, patterns and profiles
  - post-translational modification
  - protein structure
  - protein-protein interaction
  - similarity search/alignment
- genomics
- structural bioinformatics
- systems biology
- phylogeny/evolution
- population genetics
- transcriptomics
- biophysics
- imaging
- IT infrastructure
- drug design

Resources A..Z

Links/Documentation

Databases

- neXtProt • human proteins • [more]
- PROSITE • protein domains and families • [more]
- STRING • protein-protein interactions • [more]
- SWISS-MODEL Repository • protein structure homology models • [more]
- UniProtKB • functional information on proteins • [more]
- UniProtKB/Swiss-Prot • protein sequence database • [more]
- ViralZone • portal to viral UniProtKB entries • [more]

- EMBnet services • bioinformatics tools, databases and courses • [more]
- ENZYME • enzyme nomenclature • [more]
- GlycoSuiteDB • glycan database • [more]
- GPSDB • gene and protein synonyms • [more]
- HAMAP • UniProtKB family classification and annotation • [more]
- MetaNetX • Metabolic Network Repository & Analysis • [more]
- MIAPEGelDB • MIAPE document edition • [more]
- MyHits • protein domains database and tools • [more]
- PANDITplus • protein families and domains resources • [more]
- PaxDb • protein abundance database • [more]
- Prolune • Popular science articles (in French) • [more]
- Protein Model Portal • structural information for a protein • [more]
- Protein Spotlight • Informally written reviews on proteins • [more]
- SugarBind • pathogen sugar-binding • [more]
- SWISS-2DPAGE • proteins on 2-D and SDS PAGE maps • [more]
- SwissSidechain • non-natural amino-acid sidechains • [more]
- SwissVar • variants in UniProtKB entries • [more]
- TCS • interaction specificity in two-component systems • [more]
- UniMES (UniProt metagenomic samples) • UniProt Metagenomic and Environmental Sequences • [more]
- UniParc (UniProt sequence archive) • UniProt sequence archive • [more]
- UniPathway • metabolic pathways for the UniProtKB • [more]
- UniRef (UniProt sequence clusters) • UniProtKB sequence clusters • [more]
- World-2DPAGE Constellation • set of 2DPAGE resources • [more]
- World-2DPAGE Repository • gel-based proteomics data • [more]

Tools

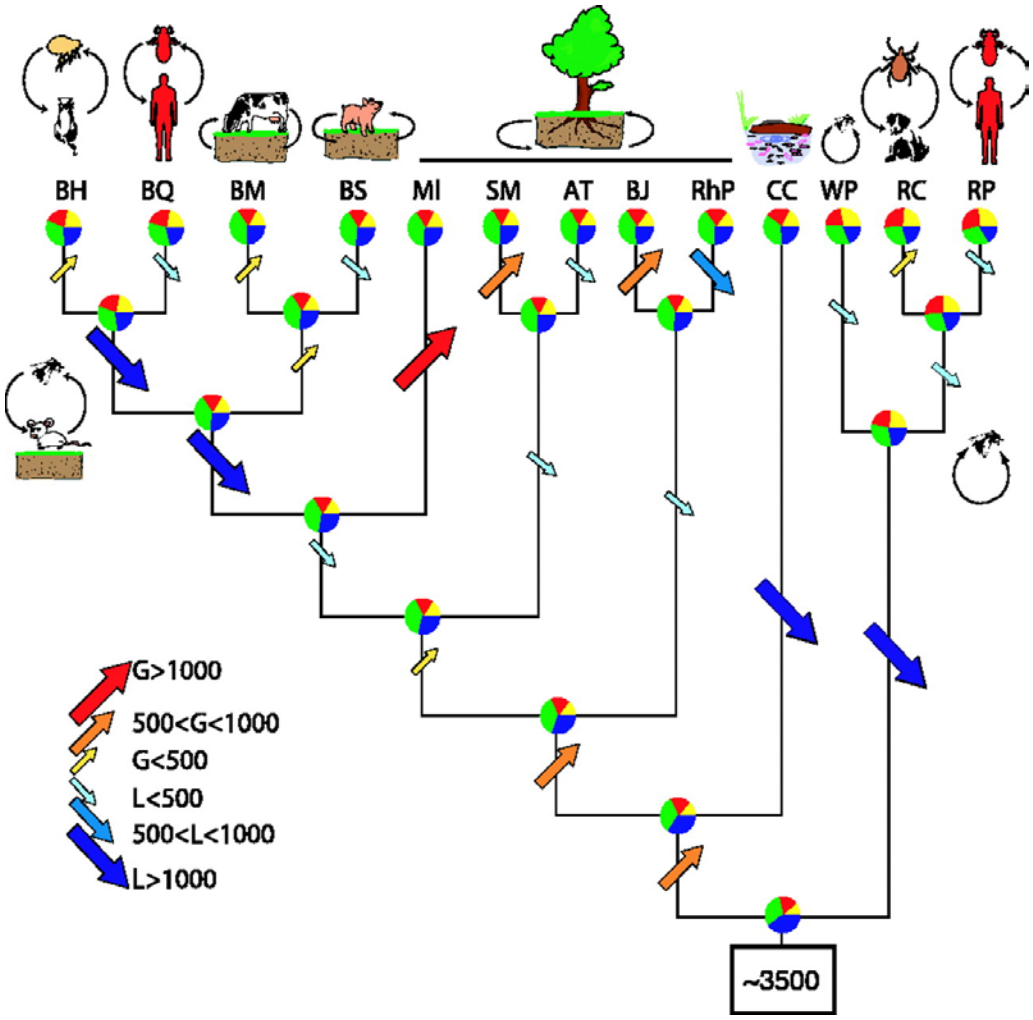
- SWISS-MODEL Workspace • structure homology-modeling • [more]
- SwissDock • protein ligand docking server • [more]

- 2ZIP • Prediction of leucine zipper domains • [more]
- 3of5 • find user-defined patterns in protein sequences • [more]
- AACompIdent • protein identification by aa composition • [more]
- AACompSim • amino acid composition comparison • [more]
- Agadir • Prediction of the helical content of peptides • [more]
- ALF • simulation of genome evolution • [more]
- Alignment tools • Four tools for multiple alignments • [more]
- AllAll • protein sequences comparisons • [more]
- APSSP • Advanced Protein Secondary Structure Prediction • [more]
- Ascalaph • Molecular modeling software • [more]
- big-PI • predict GPI modification sites • [more]
- Biochemical Pathways • Biochemical Pathways • [more]
- BLAST • sequence similarity search • [more]
- BLAST (UniProt) • BLAST search on the UniProt web site • [more]
- BLAST - NCBI • Biological sequence similarity search • [more]
- BLAST - PBIL • BLAST search on protein sequence databases • [more]
- Blast2Fasta • Blast to Fasta conversion • [more]
- boxshade • MSA pretty printer • [more]
- CFSSP • Protein secondary structure prediction • [more]
- ChloroP • chloroplast transit peptides & cleavage sites • [more]
- Click2Drug • Directory of computational drug design tools • [more]
- ClustalO (UniProt) • Align two or more protein sequences • [more]
- ClustalW • Multiple sequence alignment • [more]
- ClustalW - PBIL • Multiple sequence alignment program • [more]
- ClustalW2 • Multiple sequence alignment program • [more]
- Coiled-Coils prediction • Prediction of coiled coils regions • [more]
- COILS • Prediction of Coiled Coil Regions in Proteins • [more]
- ColorSeq • Color Protein Sequence • [more]
- Compute pI/MW • theoretical pI and Mw computation • [more]
- CPHmodels • Protein homology modeling • [more]
- CSS-Palm • Prediction of palmitoylation sites in proteins • [more]
- DAS-TMfilter • Prediction of transmembrane regions • [more]
- Decrease redundancy • Sequence redundancy reduction • [more]
- DIALIGN • Local multiple sequence aligment • [more]
- DictyOGlyc • GlcNAc O-glycosylation sites in D.discoideum • [more]
- DisEMBL • Prediction of disordered protein regions • [more]

Além de construir filogenias

- inferência de genes compartilhados e não compartilhados entre organismos de um mesmo grupo permite inferir “ganhos” e “perdas” de genes ao longo da evolução
- Exemplo: proteobactérias alfa

Fig. 4. Net gene loss or gain throughout the evolution of the {alpha}-proteobacterial species



Boussau, Bastien et al. (2004) Proc. Natl. Acad. Sci. USA 101, 9722-9727