Ferramentas de Bioinformática para Análise de Estrutura de Proteínas

ICB5731 / IBI5731

Aula 7

Ortologia e contexto genômico

Robson Francisco de Souza. Ph.D

LEEP: Laboratório de Estrutura e Evolução de Proteínas

Tópicos

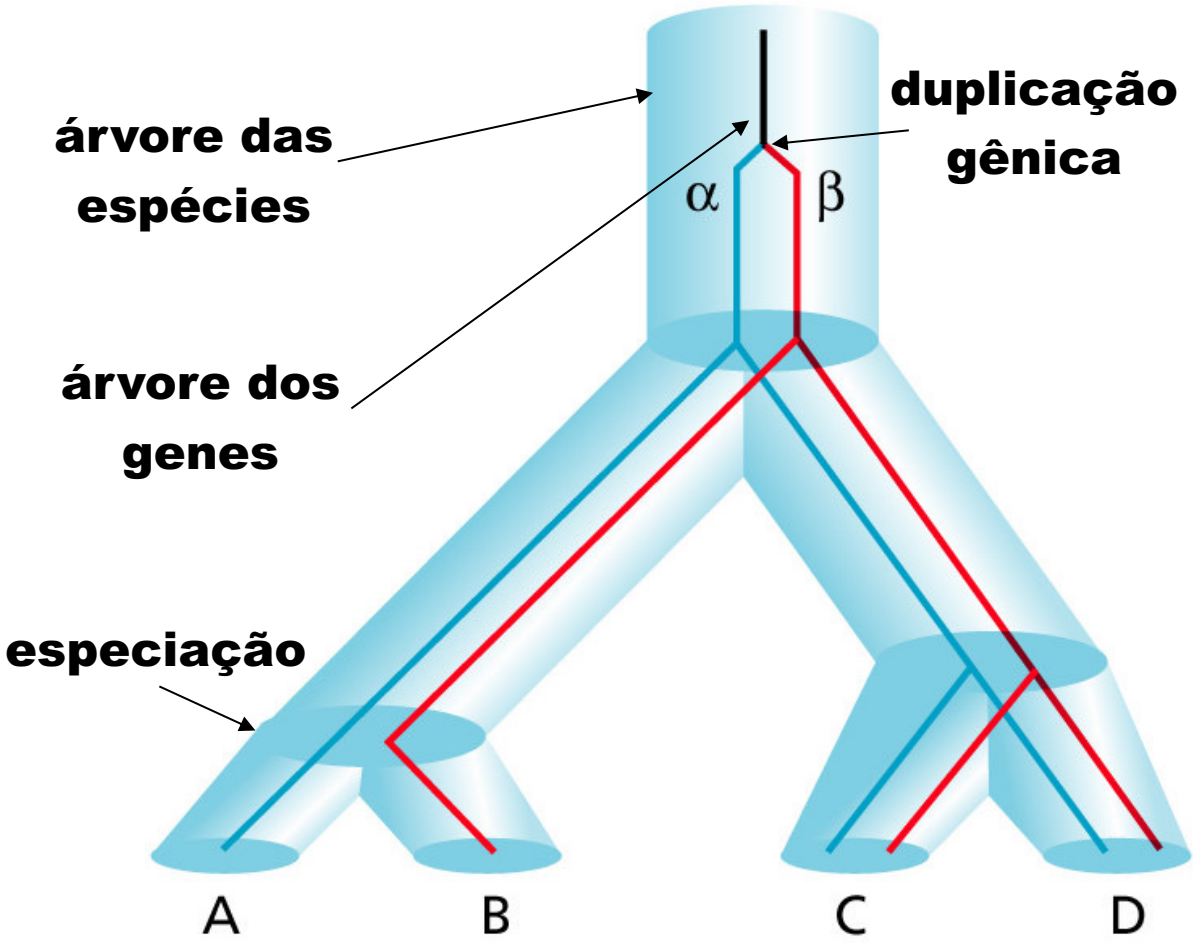
- Classificação de Ortólogos e Parálogos
  - Métodos
  - Bancos de dados
- Contexto genômico

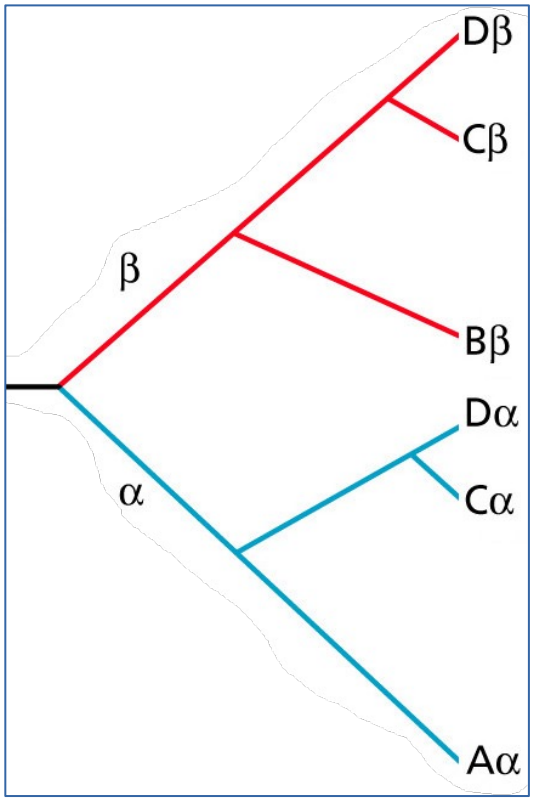
Classificação de ortólogos e parálogos

- **Problema**
  - Identificar parálogos e separá-los dos ortólogos

- **Aplicações**
  - Transferência de anotação (alta precisão!)
  - Inferência Filogenética / filogenômica
  - Predição de interações: contexto genômico

Tipos de homólogos





Árvore dos genes

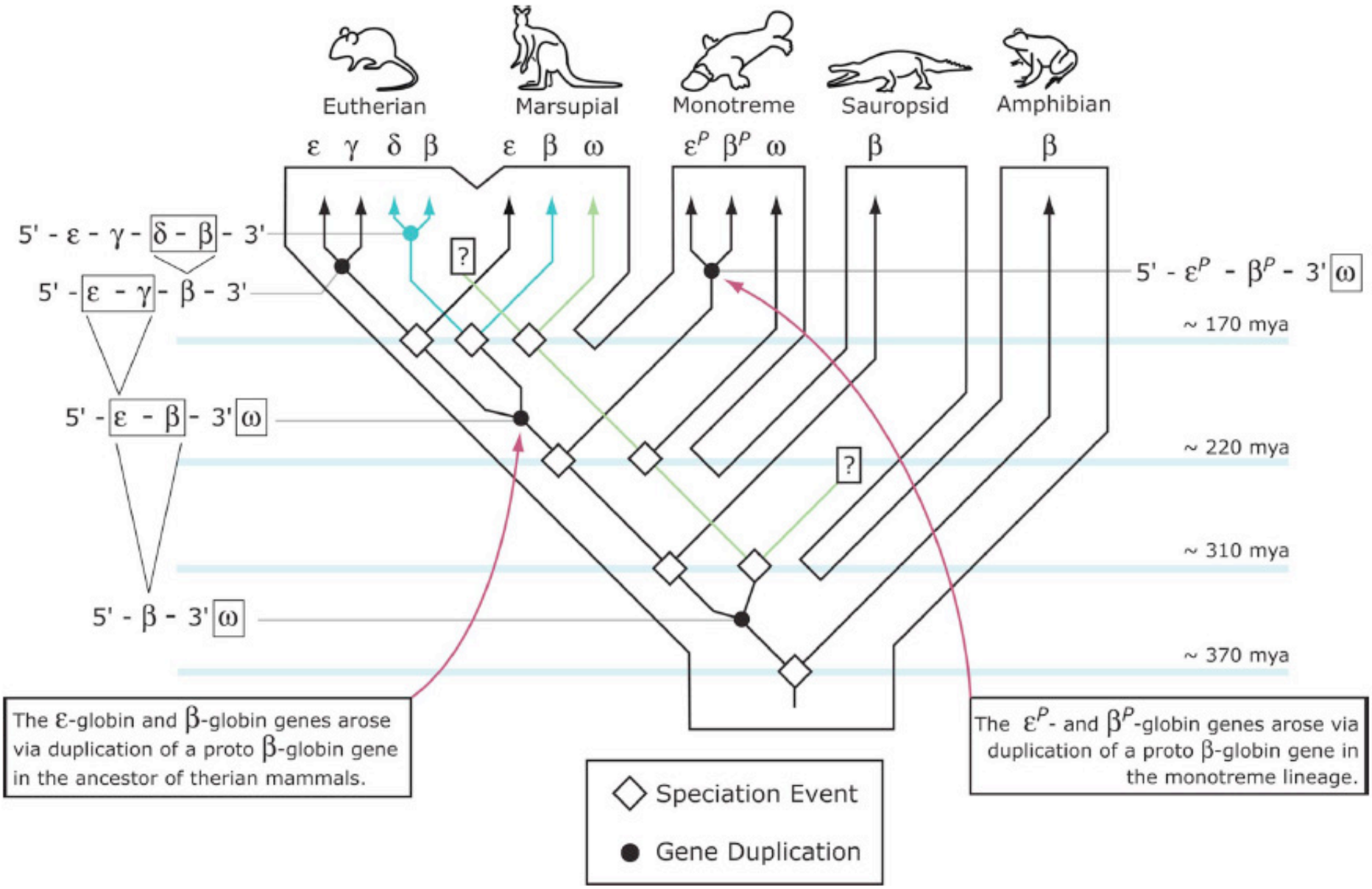
Aα, Cα e Dα : ortólogos

Bβ, Cβ e Dβ : ortólogos

qualquer α e qualquer β : parálogos

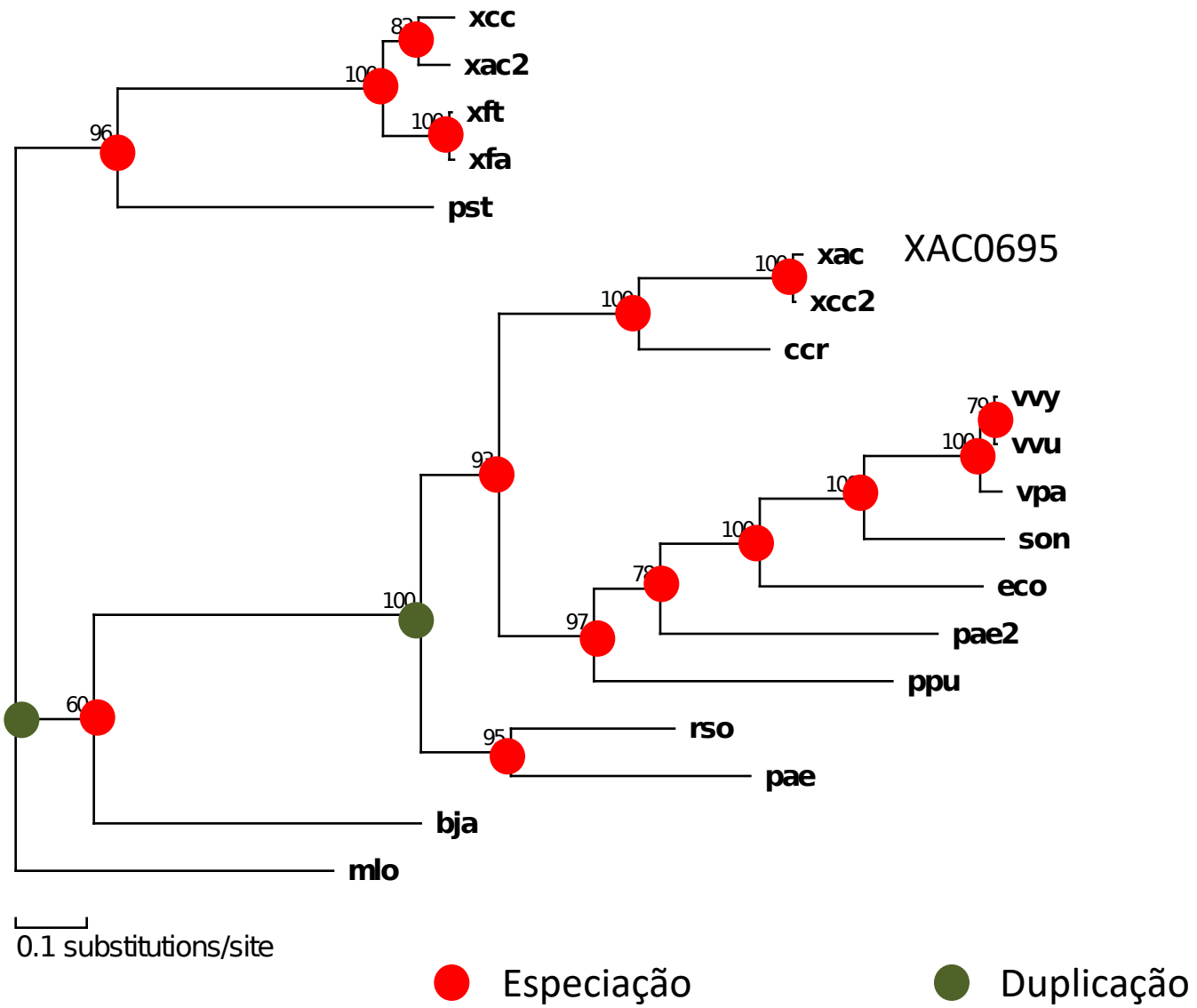
Adaptado de slide por *J. M. P. Alves*

Exemplo real: evolução das globinas



Adaptado de slide por *J. M. P. Alves*

Parálogos e ortólogos: xcsD



Parálogos e ortólogos: métodos

- **Baseados em árvores**

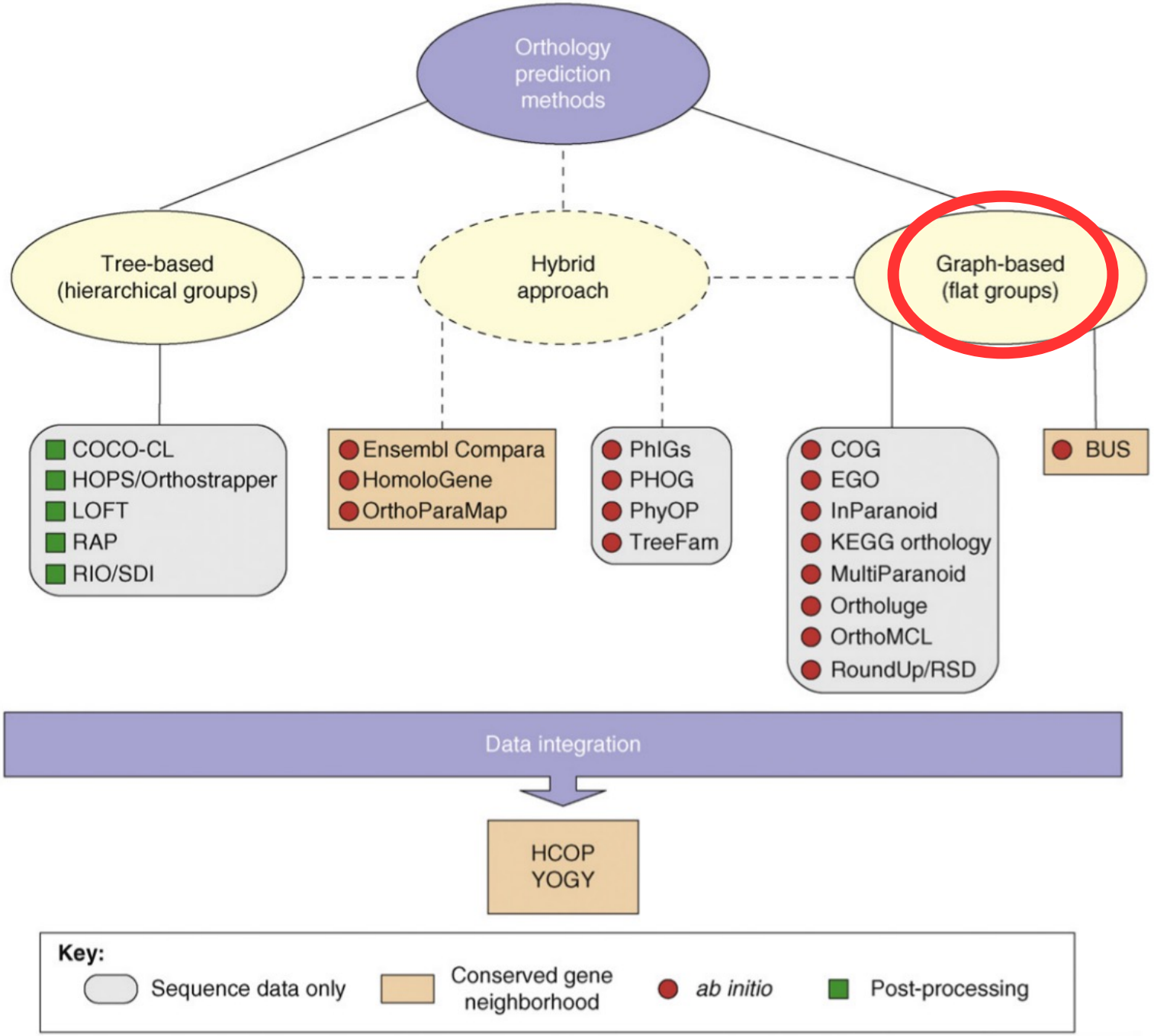
  Mais confiável se feito corretamente, porém… na maioria das vezes, é (i) dispendioso computacionalmente e (ii) difícil de automatizar sem erros

- **Baseados em grafos (similaridade)**

  Geralmente rápidos, mais sujeitos a erros (por ex.: extinção diferencial de genes)

- **Híbridos**

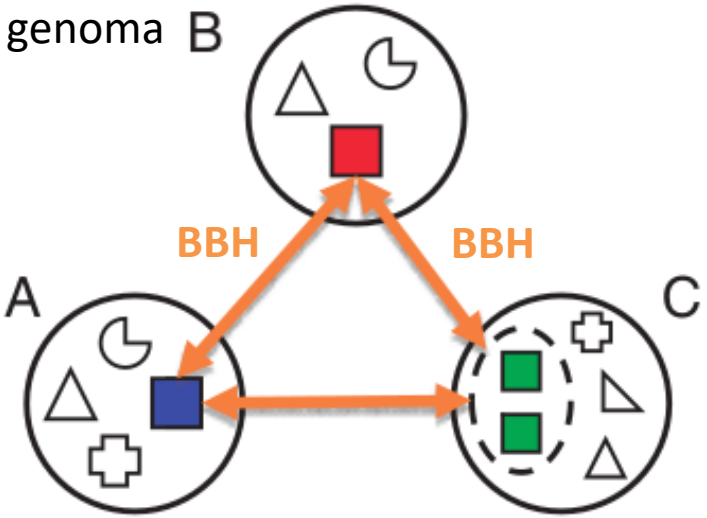
  Combinam aspectos dos dois anteriores, ainda incipientes



Kuzniar et al. (2008) Trends in Genetics, 24(11): 539–551

Métodos baseados em grafos

1. Calcular similaridade de **todas** as proteínas de cada genoma **contra todas** as proteínas de todos os outros genomas
2. Construir um **grafo** (rede) relacionando todos os resultados (similaridade)
3. Escolher os melhores resultados recíprocos (BBH: *best bidirectional hit*) ou filtrar com outro critério
4. Refinar e estender os grupamentos presentes no grafo (vários métodos: *single-linkage*, *complete-linkage*, Markov *clustering*, etc.)
5. A classificação se sustenta mesmo se o nível absoluto de similaridade entre as sequências de proteínas em questão for relativamente baixo.

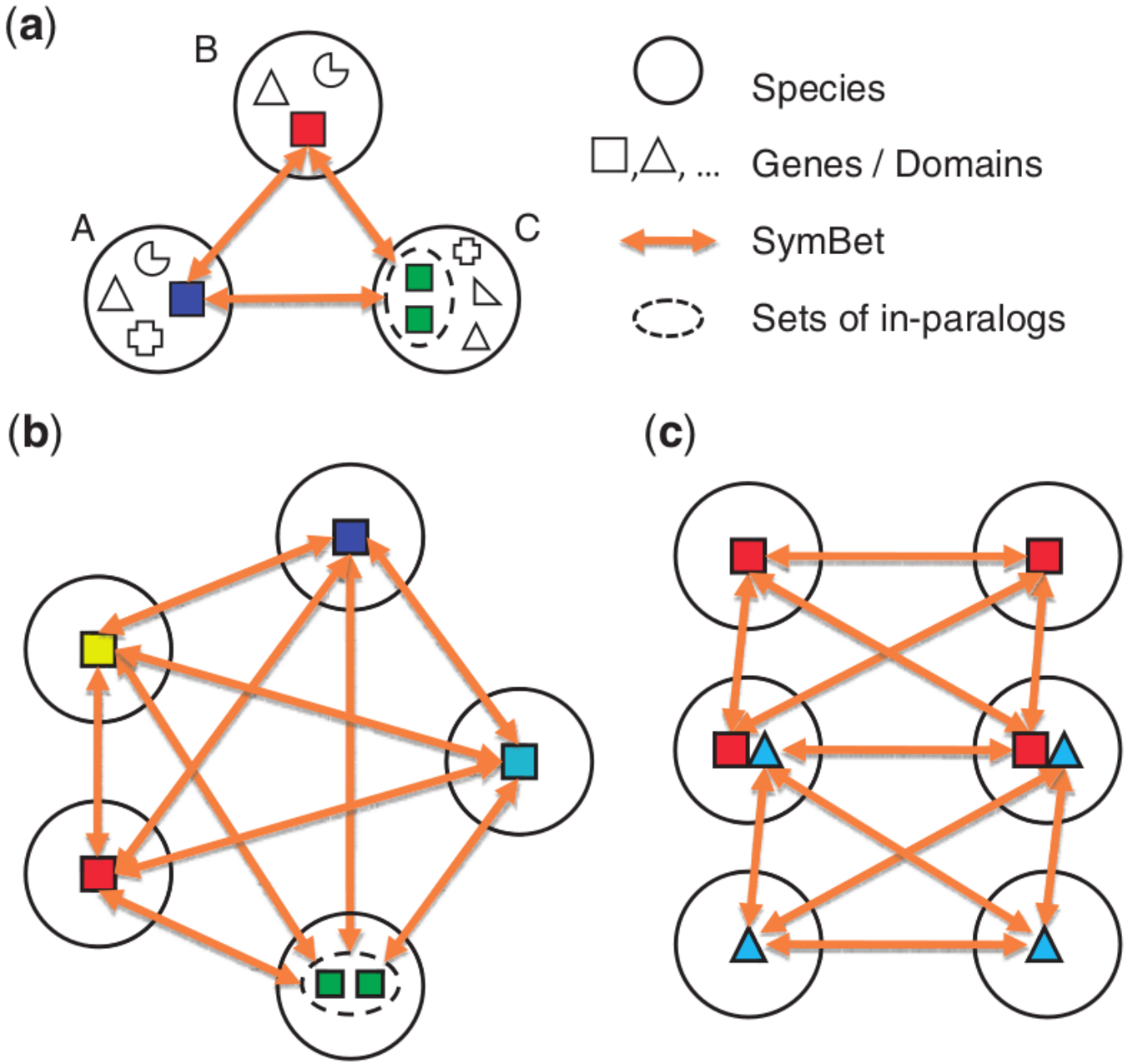


Cluster of Orthologous Groups: método

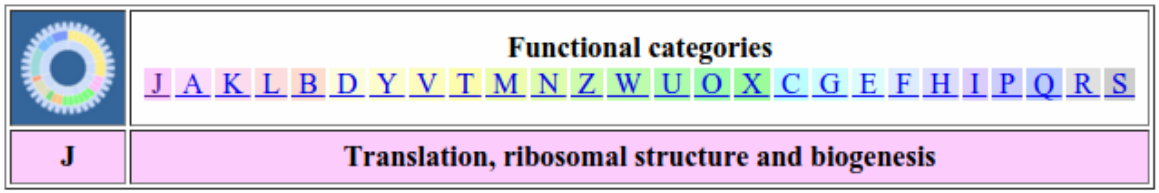
1. Fazer uma comparação das sequências proteicas todas-contra-todas.
2. Detectar e colapsar *in-parálogos*, isto é, proteínas do mesmo genoma que são mais similares entre si do que a quaisquer proteínas de outras espécies.
3. Detectar triângulos de BBHs mutuamente consistentes (*genome-specific best hits - BeTs*), levando em consideração os parálogos identificados.
4. Juntar os triângulos com um lado comum para formar COGs.
5. Realizar uma análise caso a caso (**manualmente!**) de cada COG.
   - Eliminar falsos positivos
   - Identificar proteínas de múltiplos domínios através da representação gráfica do BLAST.
   - Separar essas sequências em segmentos de domínios únicos e repetir etapas 1-4, de forma a classificar domínios individuais de acordo com suas afinidades evolutivas.
6. Examinar os grandes COGs com técnicas filogenéticas de forma a incluir novos membros ou separar alguns COGs em grupos menores.

Fonte: The NCBI Handbook - http://www.ncbi.nlm.nih.gov/books/NBK21101/

Cluster of Orthologous Groups: método



http://www.ncbi.nlm.nih.gov/COG/



Functional categories

J A K L B D Y V T M N Z W U O X C G E F H I P Q R S

J — **Translation, ribosomal structure and biogenesis**

org	pro	COG	cat	annotation
709	1356	COG0008	J	Glutamyl- or glutaminyl-tRNA synthetase
697	1021	COG0009	J	tRNA A37 threonylcarbamoyladenosine synthetase subunit TsaC/SUA5/YrdC
699	728	COG0012	J	Ribosome-binding ATPase YchF, GTP1/OBG family
708	812	COG0013	J	Alanyl-tRNA synthetase
707	731	COG0016	J	Phenylalanyl-tRNA synthetase alpha subunit
410	488	COG0017	J	Aspartyl/asparaginyl-tRNA synthetase
700	763	COG0018	J	Arginyl-tRNA synthetase
261	275	COG0023	J	Translation initiation factor 1 (eIF-1/SUI1)
703	983	COG0024	J	Methionine aminopeptidase
702	825	COG0030	J	16S rRNA A1518 and A1519 N6-dimethyltransferase RsmA/KsgA/DIM1 (may also have DNA glycosylase/AP lyase activity)
703	1150	COG0037	J	tRNA(Ile)-lysidine synthase TilS/MesJ
619	972	COG0042	J	tRNA-dihydrouridine synthase

Outros bancos de dados de ortólogos

- **COG**

  Clusters of Orthologous Genes

- **KEGG** (KEGG Orthologs)

  Kyoto Encyclopedia of Genes and Genomes

- **eggNOG (http://eggnogdb.embl.de/app/home)**

  Evolutionary genealogy of genes: non-supervised Orthologous Groups

- **Outros**

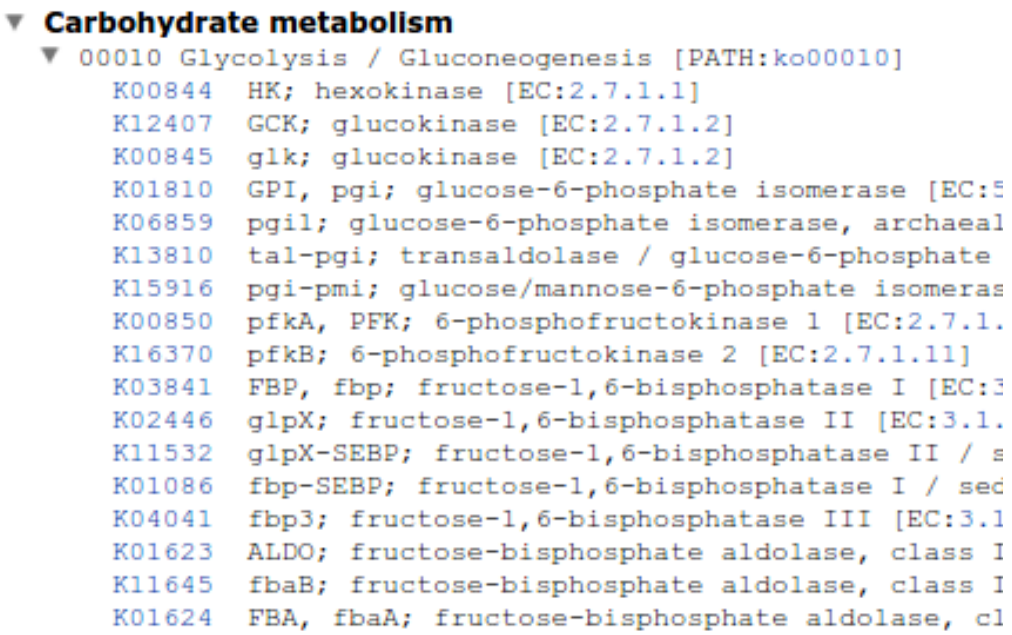
  OrthoMCL, InParanoid, etc. etc.

KEGG

- Kyoto Encyclopedia of Genes and Genomes (1995)
- Coleção de 17 bancos de dados diferentes
- O principal: vias metabólicas
- Inclui classificação de ortólogos própria (KO)

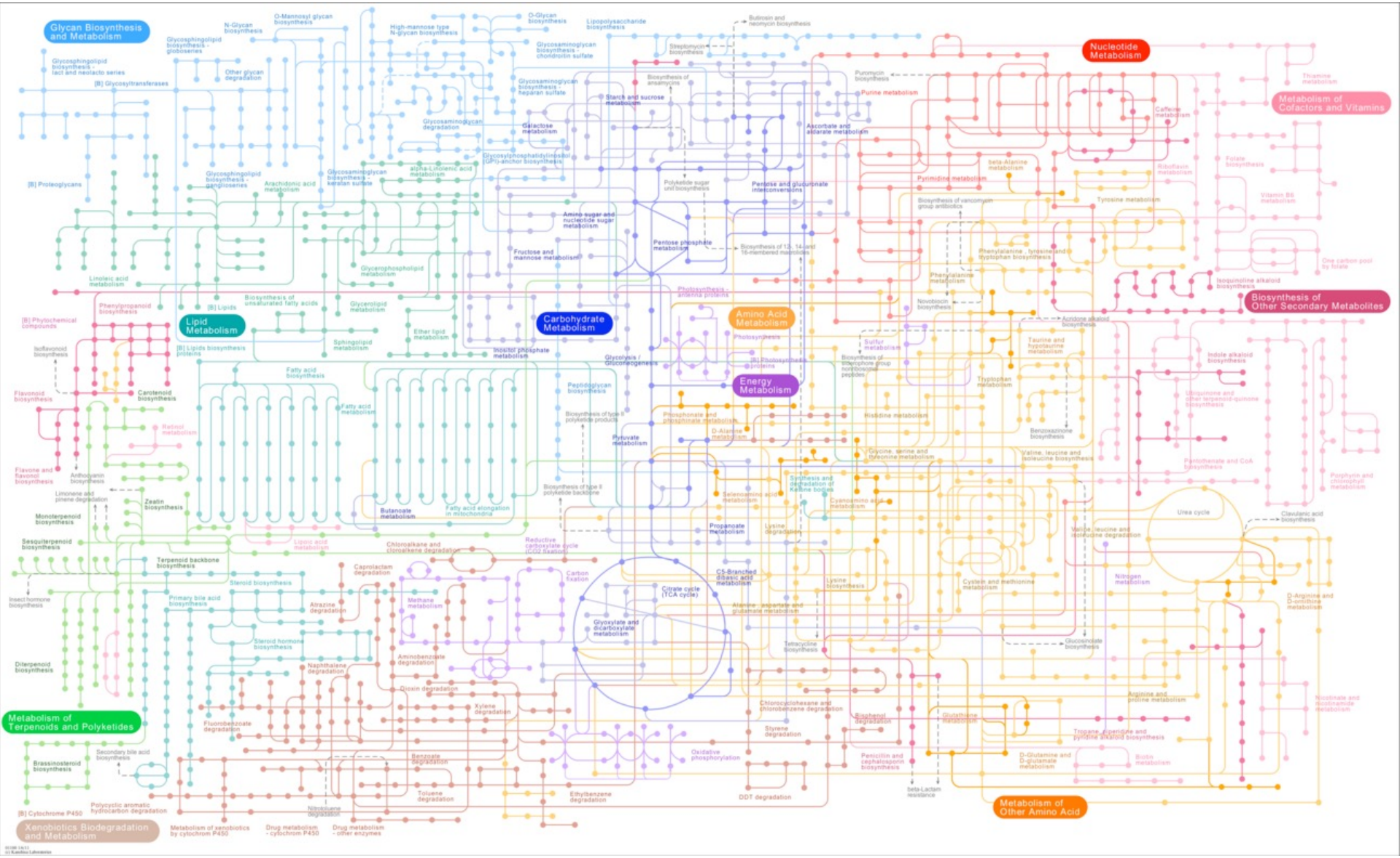
Category	Database	Content
Systems information	KEGG PATHWAY	KEGG pathway maps
	KEGG BRITE	BRITE functional hierarchies
	KEGG MODULE	KEGG modules of functional units
Genomic information	KEGG ORTHOLOGY	KEGG Orthology (KO) groups
	KEGG GENOME	KEGG organisms with complete genomes
	KEGG GENES	Gene catalogs of complete genomes
	KEGG SSDB	Sequence similarity database for GENES
Chemical information	KEGG COMPOUND	Metabolites and other small molecules
	KEGG GLYCAN	Glycans
	KEGG REACTION	Biochemical reactions
	KEGG RPAIR	Reactant pair chemical transformations
	KEGG RCLASS	Reaction class defined by RPAIR
	KEGG ENZYME	Enzyme nomenclature
Health information	KEGG DISEASE	Human diseases
	KEGG DRUG	Drugs
	KEGG DGROUP	Drug groups
	KEGG ENVIRON	Crude drugs and health-related substances

Chemical information category is collectively called KEGG LIGAND
Health information category is collectively called KEGG MEDICUS

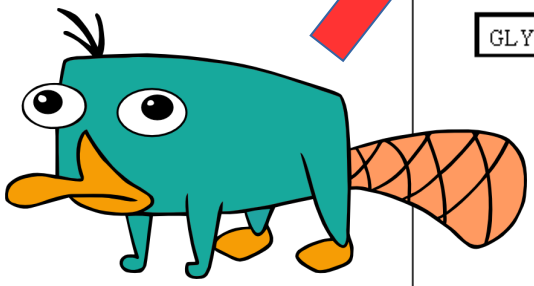


http://www.genome.jp/kegg/

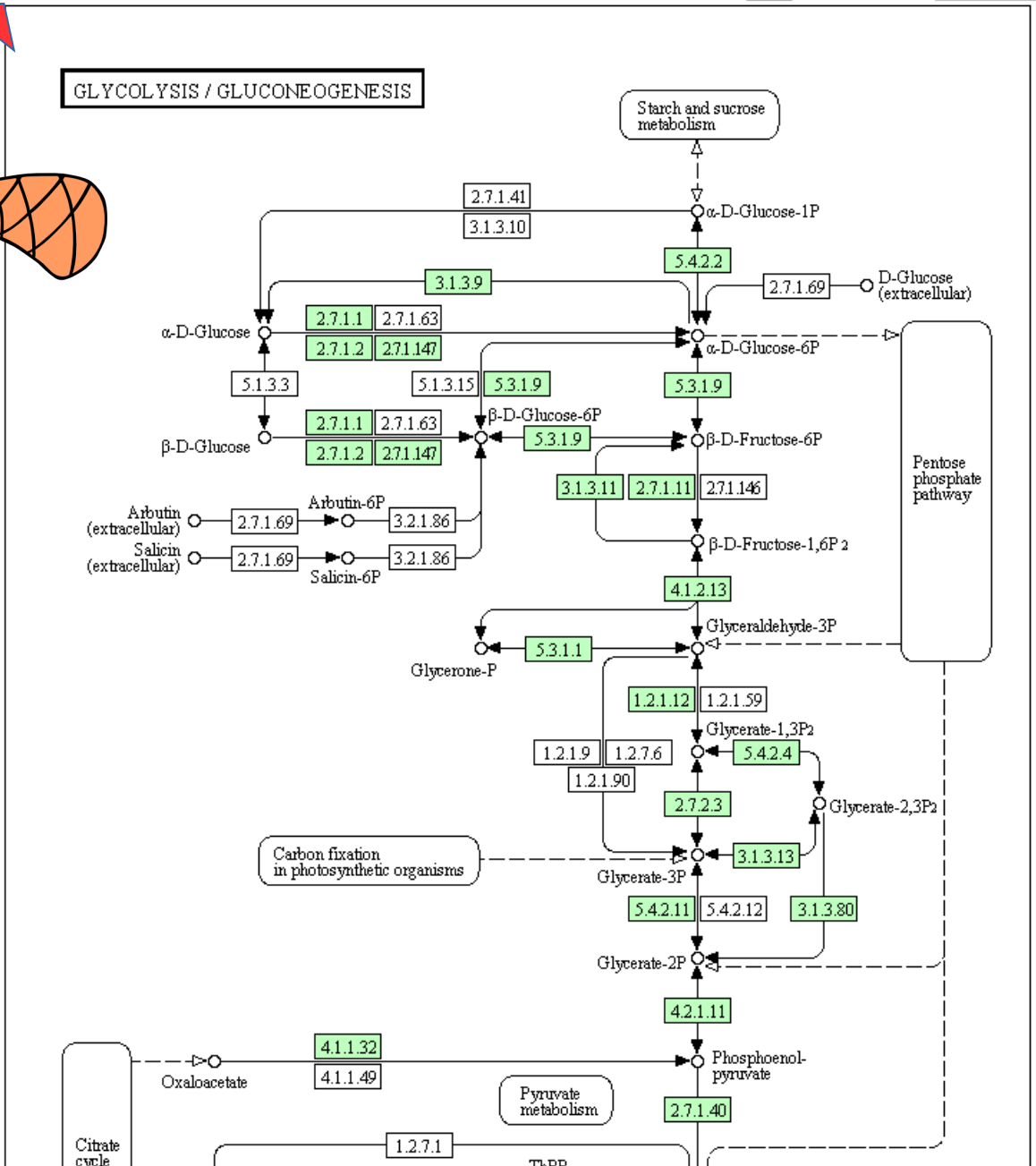
Vias metabólicas de fato



Ornitorrinco faz glicólise!?



Perry the Platypus!



J. M. P. Alves

Vias

(das quais essa proteína participa)

KEGG **Ornithorhynchus anatinus (platypus): 100079206** Help

Entry	100079206 CDS T01045
Gene name	GPI
Definition	glucose-6-phosphate isomerase
KO	K01810 ucose-6-phosphate isomerase [EC:5.3.1.9]
Organism	Ornithorhynchus anatinus (platypus)
Pathway	oaa00010 Glycolysis / Gluconeogenesis oaa00030 Pentose phosphate pathway oaa00500 Starch and sucrose metabolism oaa00520 Amino sugar and nucleotide sugar metabolism oaa01100 Metabolic pathways oaa01130 Biosynthesis of antibiotics oaa01200 Carbon metabolism
Module	oaa_M00001 Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate oaa_M00004 Pentose phosphate pathway (Pentose phosphate cycle)
Brite	KEGG Orthology (KO) [BR:oaa00001] Metabolism Overview 01200 Carbon metabolism 100079206 (GPI) Carbohydrate metabolism 00010 Glycolysis / Gluconeogenesis 100079206 (GPI) 00030 Pentose phosphate pathway 100079206 (GPI) 00500 Starch and sucrose metabolism 100079206 (GPI) 00520 Amino sugar and nucleotide sugar metabolism 100079206 (GPI) Enzymes [BR:oaa01000] 5. Isomerases 5.3 Intramolecular oxidoreductases 5.3.1 Interconverting aldoses and ketoses, and related compoun 5.3.1.9 glucose-6-phosphate isomerase 100079206 (GPI) Exosome [BR:oaa04147] Exosomal proteins Exosomal proteins of haemopoietic cells (B-cell, T-cell, DC-cel 100079206 (GPI) Exosomal proteins of other body fluids (saliva and urine) 100079206 (GPI) Exosomal proteins of colorectal cancer cells 100079206 (GPI) BRITE hierarchy
SSDB	Ortholog Paralog GFIT
Motif	Pfam: PGI Motif
Other DBs	NCBI-ProteinID: XP_007669717 NCBI-GI: 620979733 NCBI-GeneID: 100079206
LinkDB	All DBs

KEGG

ORTHOLOGY: K02355

Help

Entry	K02355 KO
Name	fusA, GFM, EFG
Definition	elongation factor G
Disease	H00891 Combined oxidative phosphorylation deficiency (COXPD)
Brite	Translation factors [BR:ko03012] Eukaryotic Type Elongation factors K02355 fusA, GFM, EFG; elongation factor G Prokaryotic Type Elongation factors K02355 fusA, GFM, EFG; elongation factor G Mitochondrial biogenesis [BR:ko03029] Mitochondrial DNA transcription, translation, and replication fac Mitochondrial transcription and translation factors Mitochondrial translation factors K02355 fusA, GFM, EFG; elongation factor G Mitochondrial quality control factors Regulator of mitochondrial biogenesis Ubiquitas transcription factors K02355 fusA, GFM, EFG; elongation factor G BRITE hierarchy
Other DBs	COG: COG0480 GO: 0003746
Genes	HSA: 84340(GFM2) 85476(GFM1) PTR: 460812(GFM1) 471518(GFM2) PPS: 100981523(GFM1) 100991789(GFM2) GGO: 101133759 101135526 101147436(GFM1) PON: 100173642(GFM2) 100173763(GFM1) NLE: 100594248(GFM2) 100600876(GFM1) MCC: 704247(GFM2) 706456(GFM1) MCF: 102122671(GFM2) 102140197(GFM1) CSAB: 103223136(GFM2) 103241412(GFM1) RRO: 104662312(GFM2) 104665852(GFM1) » show all Taxonomy KOALA UniProt
Reference	PMID:18213444
Authors	Noble CG, Song H

All links

- Ontology (4)
  - KEGG BRITE (2)
  - GO (1)
  - COG (1)
- Disease (1)
  - KEGG DISEASE (1)
- Gene (202118)
  - KEGG GENES (6109)
  - KEGG DGENES (20)
  - KEGG MGENES (164444)
  - RefGene (30996)
  - EGENES (384)
  - OC (165)
- Protein sequence (5011)
  - UniProt (4178)
  - SWISS-PROT (833)
- Literature (3)
  - PubMed (3)
- All databases (207137)

Download RDF

Links para outros bancos

Links para genes

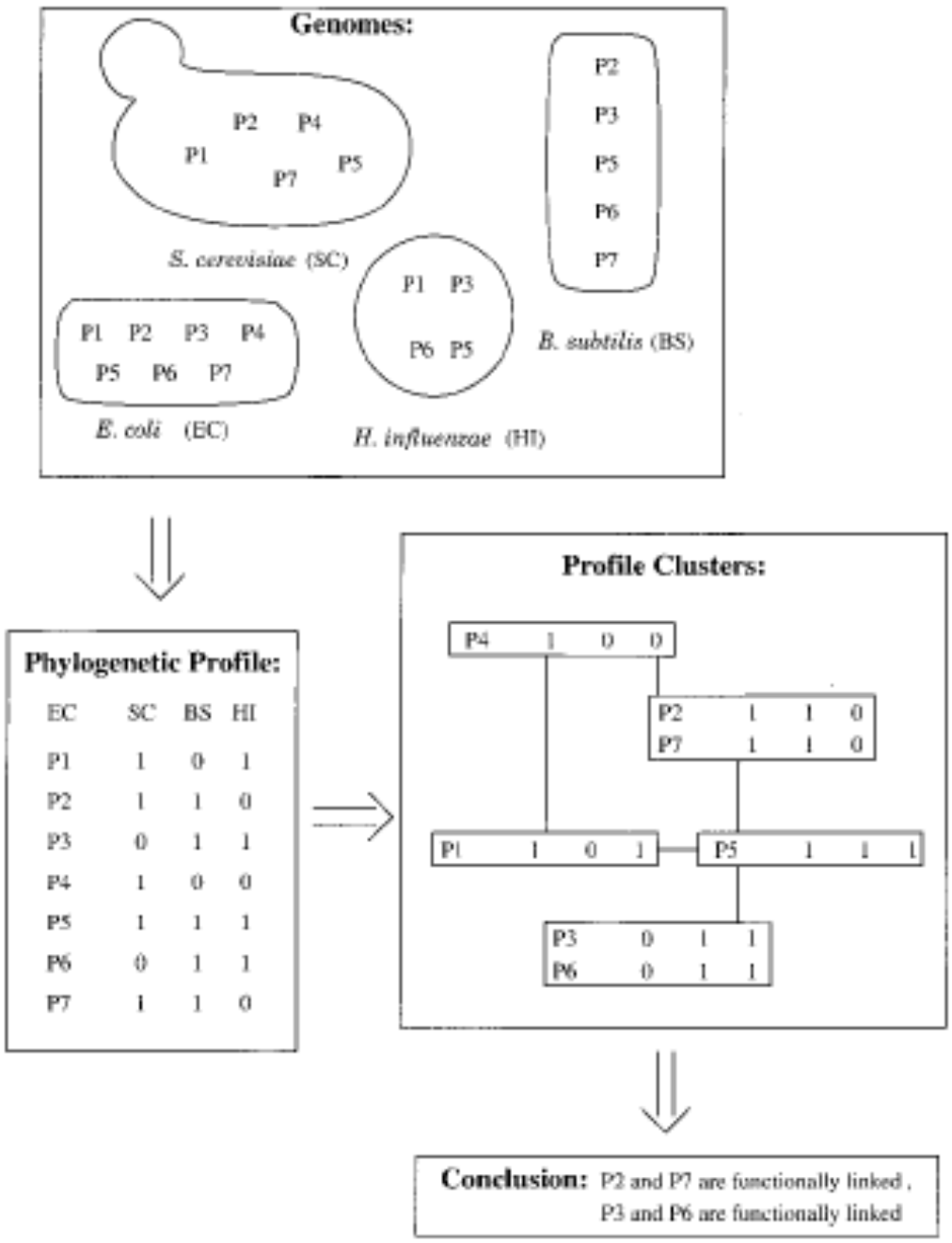
Contexto genômico

O que fazer quando nada é conhecido sobre os homólogos de um gene/proteína de interesse?

- Métodos computacionais para previsão de **vínculos funcionais**
- **Premissa**: a existência de vínculos funcionais gera correlações entre na distribuição e nas posições de genes em genomas
- Identificados usando a **comparação de genomas**
- A homologia serve apenas para identificar os genes mas não determina o padrão na sua distribuição

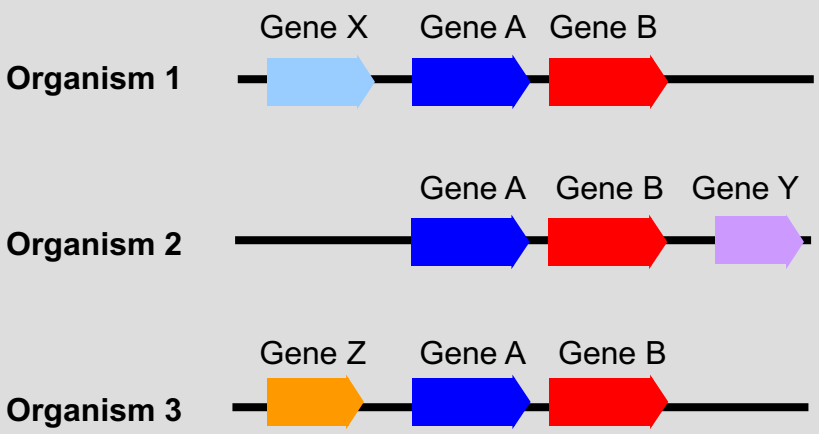
Perfil filogenético

- Correponde a uma correlação na distribuição de genes homólogos entre genomas
- Padrão de presença/ausência nos genomas
- Análogo a experimentos de knock-out



Biological Rosetta stones

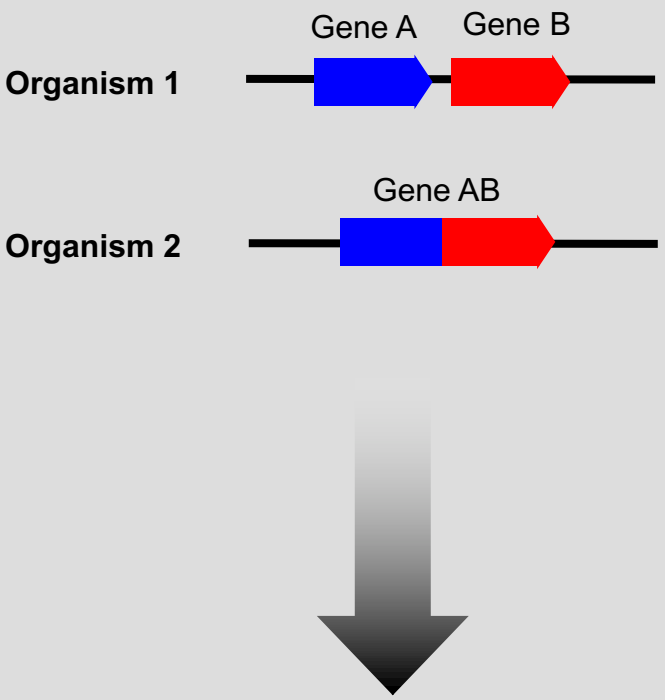
Gene neighborhoods



Inference: Selective pressure to regulate these genes together; functionally or physically linked.

Co-expression or physical interaction

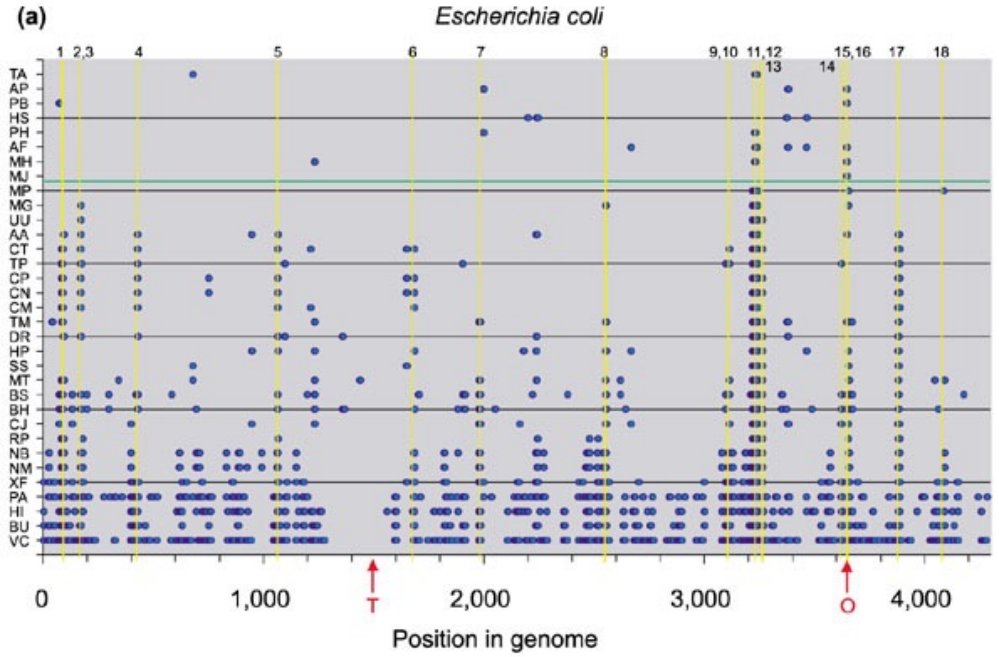
Gene fusion events

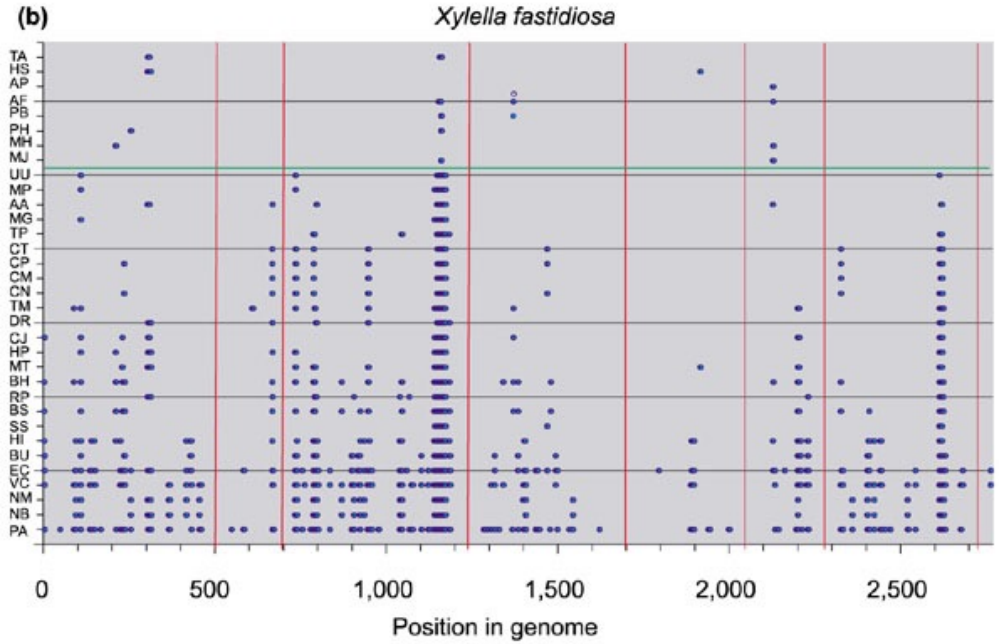


Inference: Gene fusion confers selective advantage; the two gene products are physically or functionally linked.

Protein-protein interaction or common pathway

Evolução da distribuição e ordem dos genes

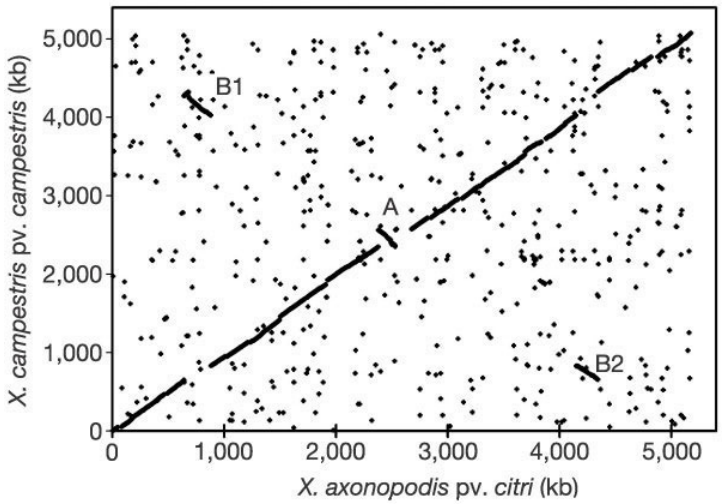




A ordem dos genes é rapidamente alterada por eventos como:

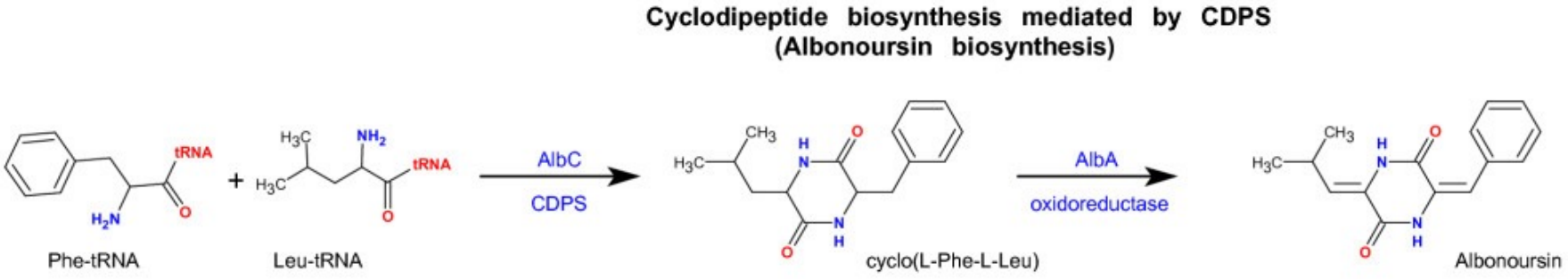
- Transposição (mudar de lugar)
- Perda de genes
- Ganho de genes (HGT)
- Rearranjos cromossômicos

mas é preservada em pequena escala



Exemplo: parálogos de tRNA sintetase

- AlbC catalisa a síntese de dipeptídeos cíclicos

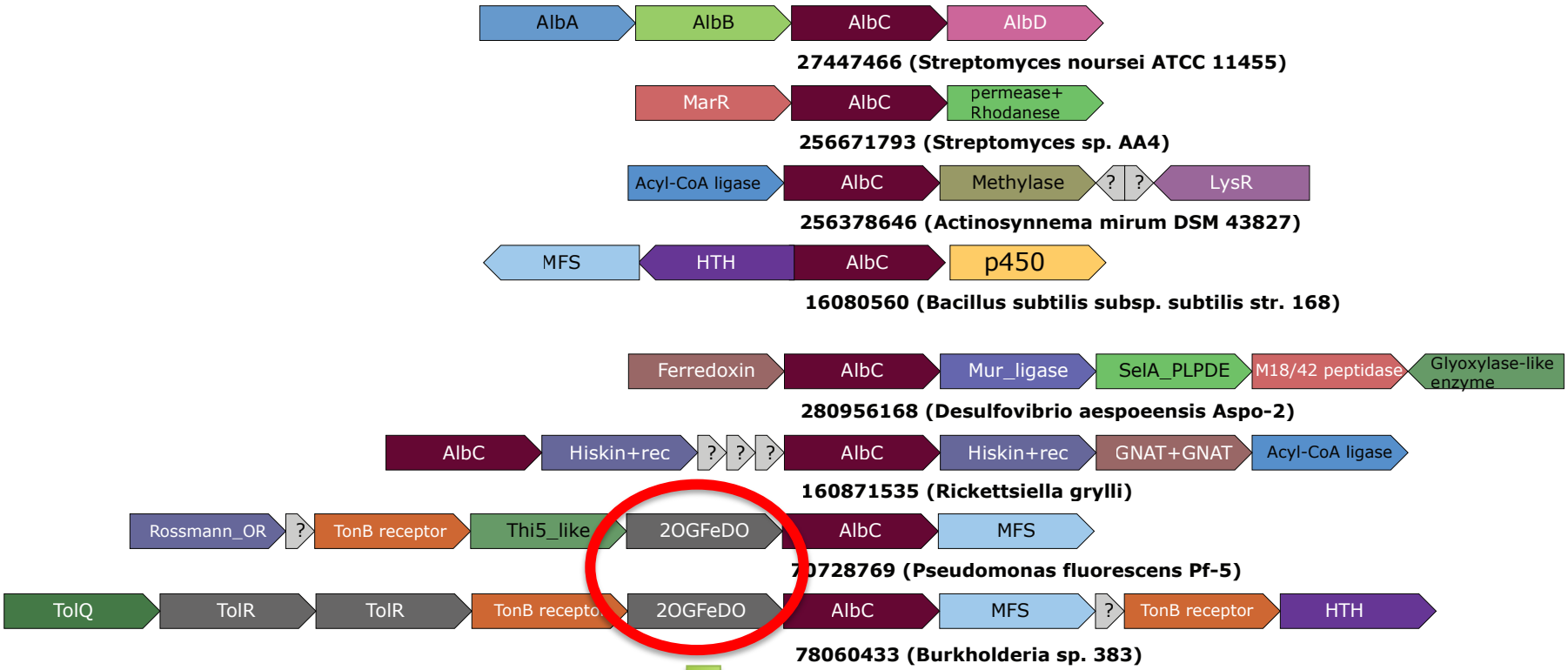


Gondry *et al*. Cyclodipeptide synthases are a family of tRNA-dependent peptide bond–forming enzymes. *Nature Chemical Biology* 5, 414 - 420 (2009)

- Descobrimos que AlbC é um parálogo remoto das tRNA sintetases da Classe-I (e.g. Tyr-tRNA syntetase) que se especializou em catalizar a formação das atua na síntese de peptídeos não ribossomais

Inferências baseadas em contexto

- Contexto dos homólogos de AlbC

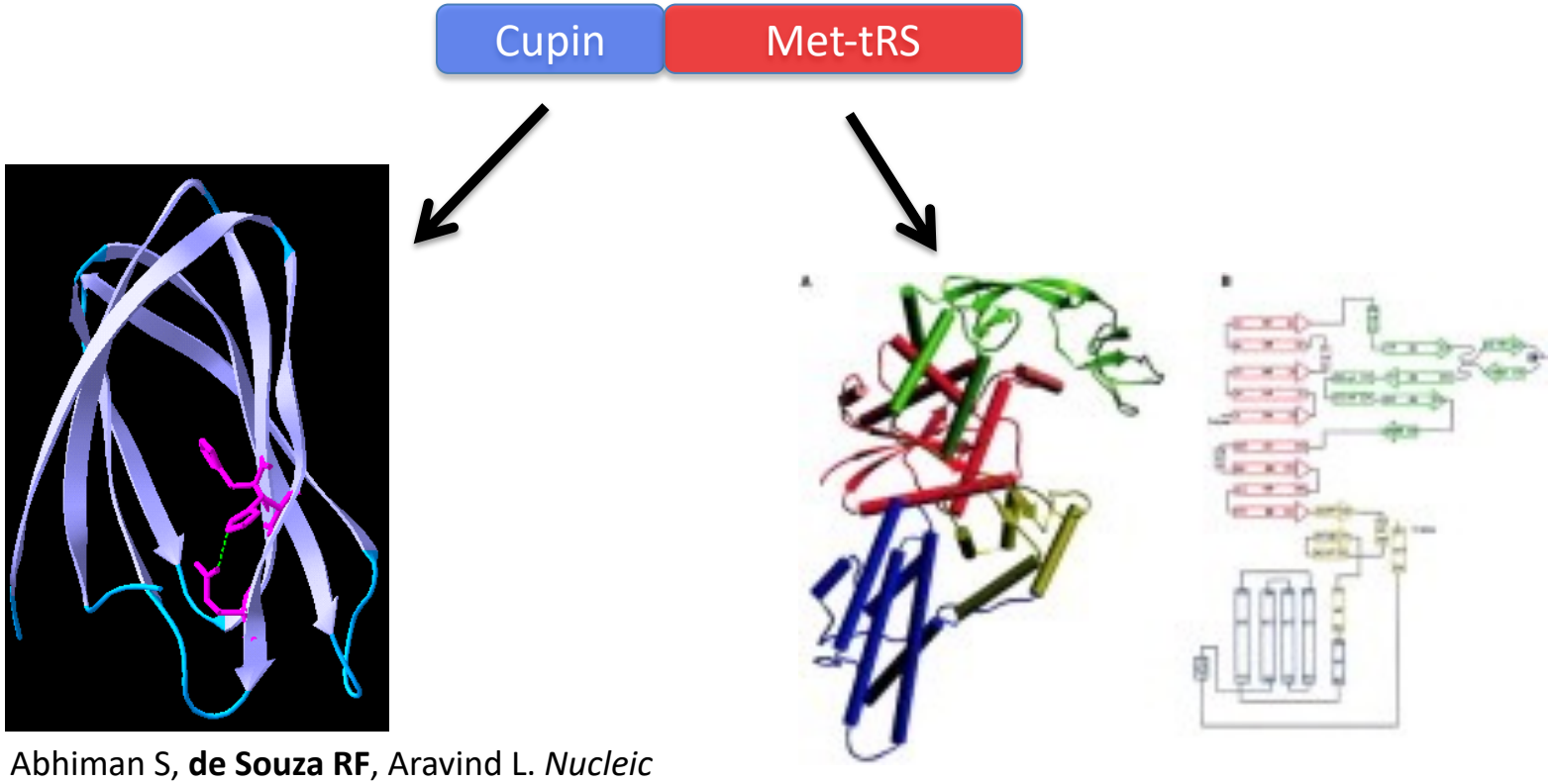


2-oxoglutarato Fe-dioxigenases:

Cupin-like dioxigenases

Inferências baseadas em contexto

Opa! Já havíamos observado fusões do domínio catalítico da metionil-tRNA syntetase com as “metal-binding cupins” (dioxigenases da classe estrutural, a.k.a fold, DSBH)



Iyer LM, Abhiman S, **de Souza RF**, Aravind L. *Nucleic Acids Research,* 38(16):5261-79 *(*2010)

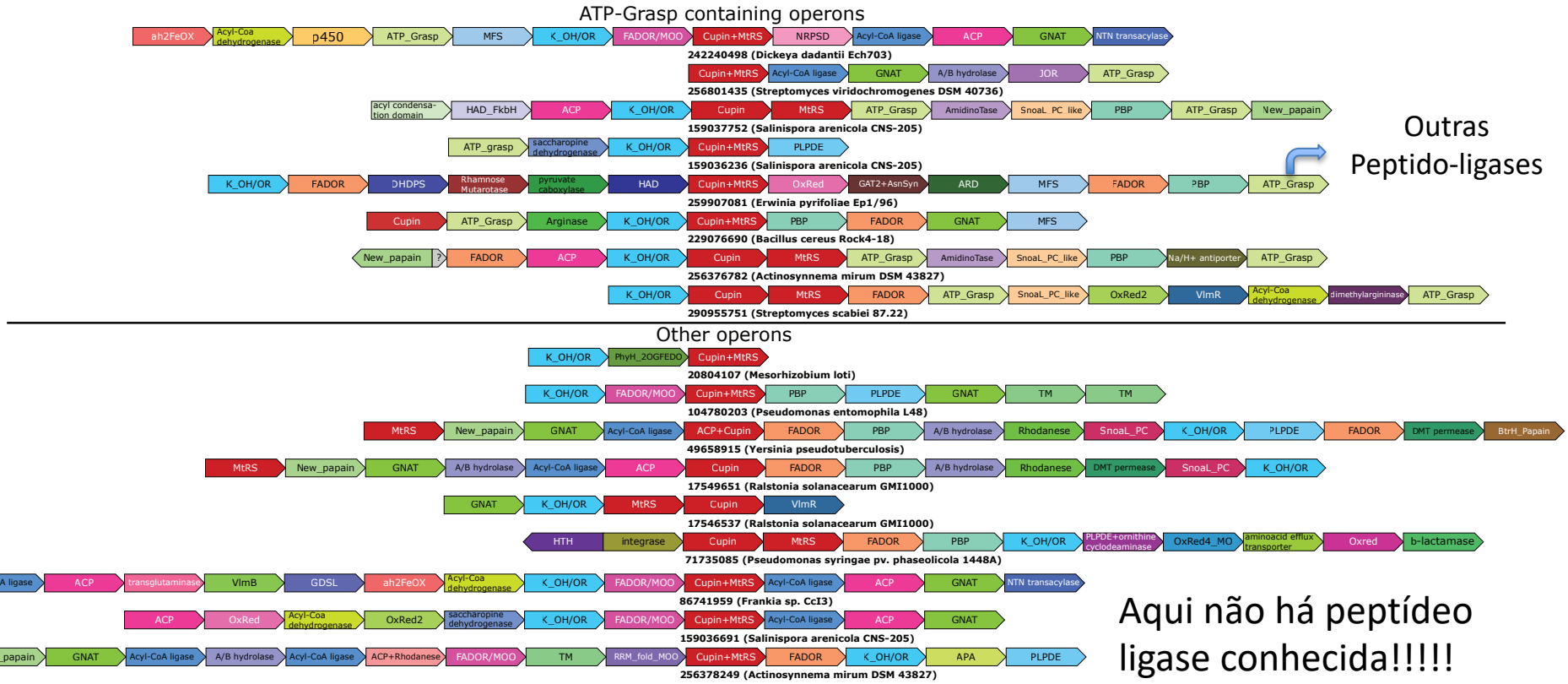
Journal of Molecular Biology 294(5) :1287-1297 (1999)

Operons de Cupin+MtRS: função

A conservação da vizinhança em diversas linhagens de bactérias fornece suporte ao vínculo com uma N6-lisina hidroxilase

Núcleo conservado: Lisina oxidoreductase + cupin + MtRS

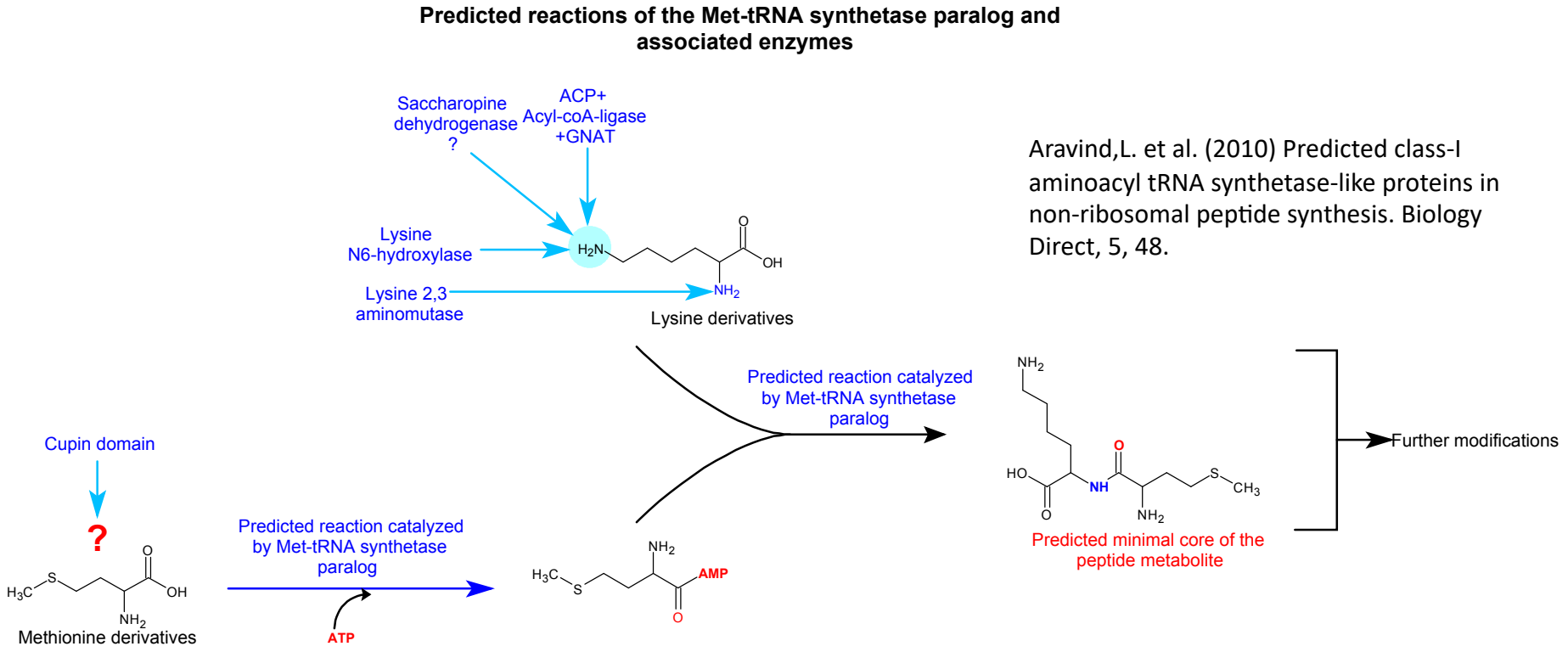
Atividades enzimaticas para dois aminoácidos (Lys, Met): o composto é pelo menos um dipeptídeo



Inferências baseadas em contexto

Resultados

- AlbC representa um exemplo de tRNA sintetase recrutada para síntese em outra via metabólica
- O parálogo de MtRS é um novo exemplo de parálogo de tRNA sintetase que catalisa a formação de uma ligação peptídica entre uma lisina e uma metionina como evidenciado pelo contexto genômico.

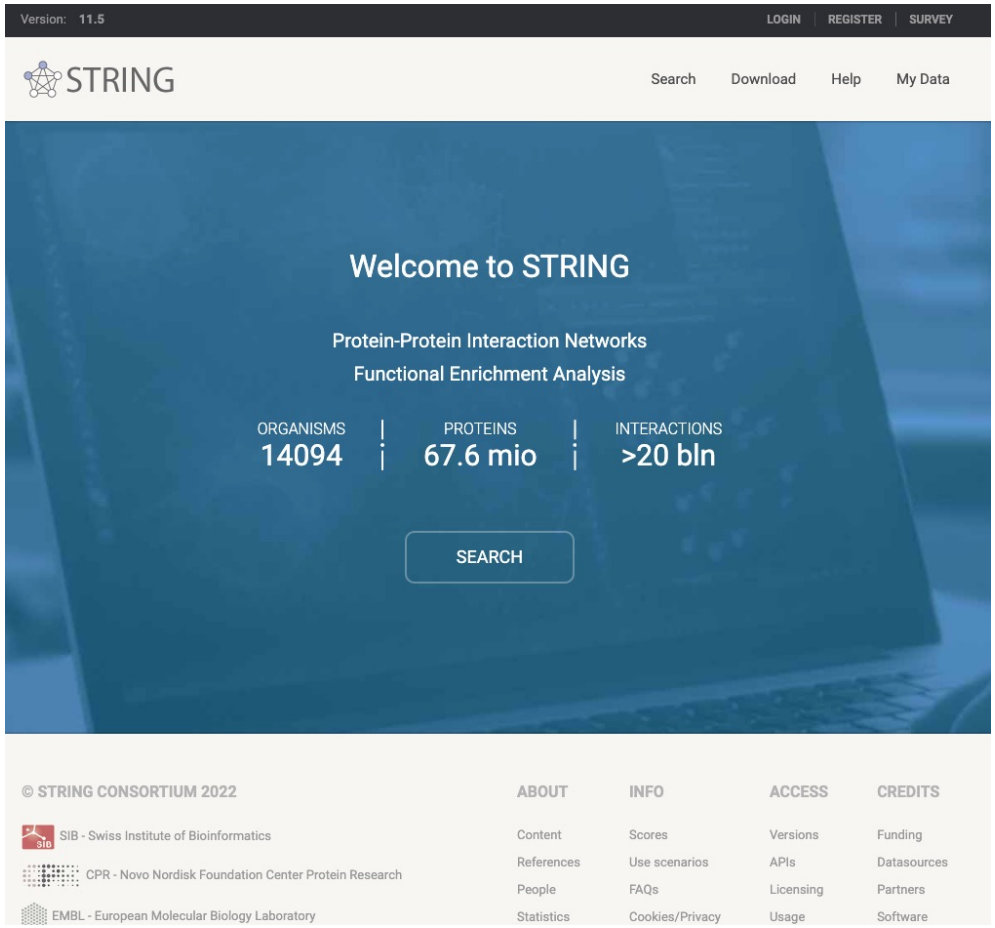


Aravind,L. et al. (2010) Predicted class-I aminoacyl tRNA synthetase-like proteins in non-ribosomal peptide synthesis. Biology Direct, 5, 48.

Banco de dados de contexto e interações

What's that gene (or protein)? Online resources for exploring functions of genes, transcripts, and proteins

James R. A. Hutchins
Institute of Human Genetics, Centre National de la Recherche Scientifique (CNRS), 34396 Montpellier, France

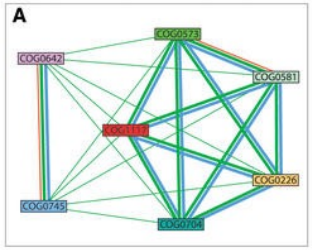


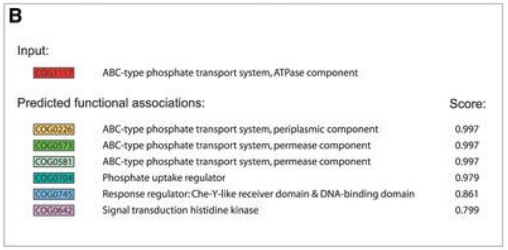
STRING

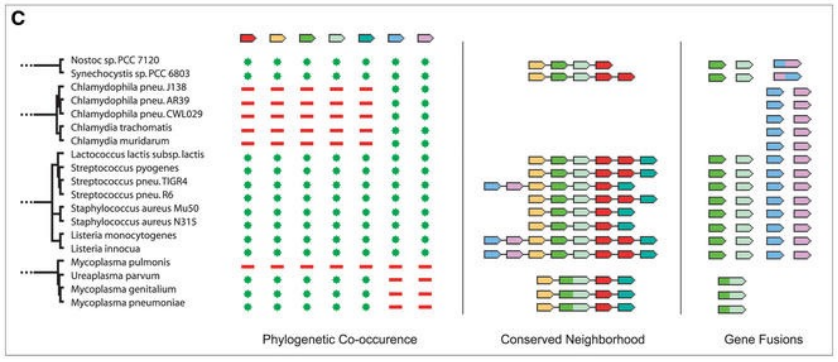
Combina múltiplos tipos de dado:

- Contexto genômico
- Expressão gênica
- Interações (PPI)
- Text mining

http://www.string-db.org







Referências

- Kumar, S. e M. Nei. Molecular Evolution and Phylogenetics. New York: Oxford University Press. 2000
- Vandamme et al. (2009)The phylogenetic handbook: A Pratical Approach to Phylogenetic Analysis and Hypothesis Testing
- Felsenstein, J. 2004. Inferring Phylogenies. Sinauer Associates, Sunderland, Massachusetts.
- Felsenstein, J (1988) Phylogenies from molecular sequences: inference and reliability. Annual Review of Genetics 22: 521-565
- Russel DJ. (2014) Multiple Sequence Alignment Methods. Springer, ISBN 978-1-62703-646-7.
- Durbin,R. et al. (1998) Biological Sequence Analysis Probabilistic Models of Proteins and Nucleic Acids 1st ed. Cambridge University Press.
- Schuster-Böckler,B. and Bateman,A. (2007) An introduction to hidden Markov models. Curr. Protoc. Bioinformatics, Appendix 3, Appendix 3A.
- Hutchins,J.R.A. (2014) What's that gene (or protein)? Online resources for exploring functions of genes, transcripts, and proteins. Mol. Biol. Cell, 25, 1187–201.
- Cox , Michael M.; Phillips Jr., George N. (2008) **Handbook of Proteins: Structure, Function and Methods** (2 volume set), Wiley-Interscience.
- Doerks,T. et al. (2004) Functional clues for hypothetical proteins based on genomic context analysis in prokaryotes. Nucleic Acids Res., 32, 6321–6.
- Snel,B. et al. (2000) STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res., 28, 3442–4.
- Zhao,S. et al. (2013) Discovery of new enzymes and metabolic pathways by using structure and genome context. Nature, 502, 698–702.

Prática de contexto genômico

- Ver contexto no KEGG (procurar ftsA e clicar em “Clusters”)
- Acessar o banco de dados String: www.string-db.org
  - Procurar pelo gene ftsA de *E. coli* K12 MG
- Procurar por ftsA no Gene Context Tool
  - http://biocomputo.ibt.unam.mx:8080/GeConT/
- Repetir as análises para AmiC e NlpD
- Repetir as análises para xcsD