Encontro França-Brasil de Bioinformática
Universidade Estadual de Santa Cruz (UESC)
Ilhéus-BA - Brasil

Genômica Comparativa

Maria Emília M. T. Walter
Departamento de Ciência da Computação
Marcelo M. Brígido
Departamento de Biologia Celular
Universidade de Brasília

Ilhéus, 09 de novembro de 2010

Métodos computacionais para identificar genes parálogos e ortólogos

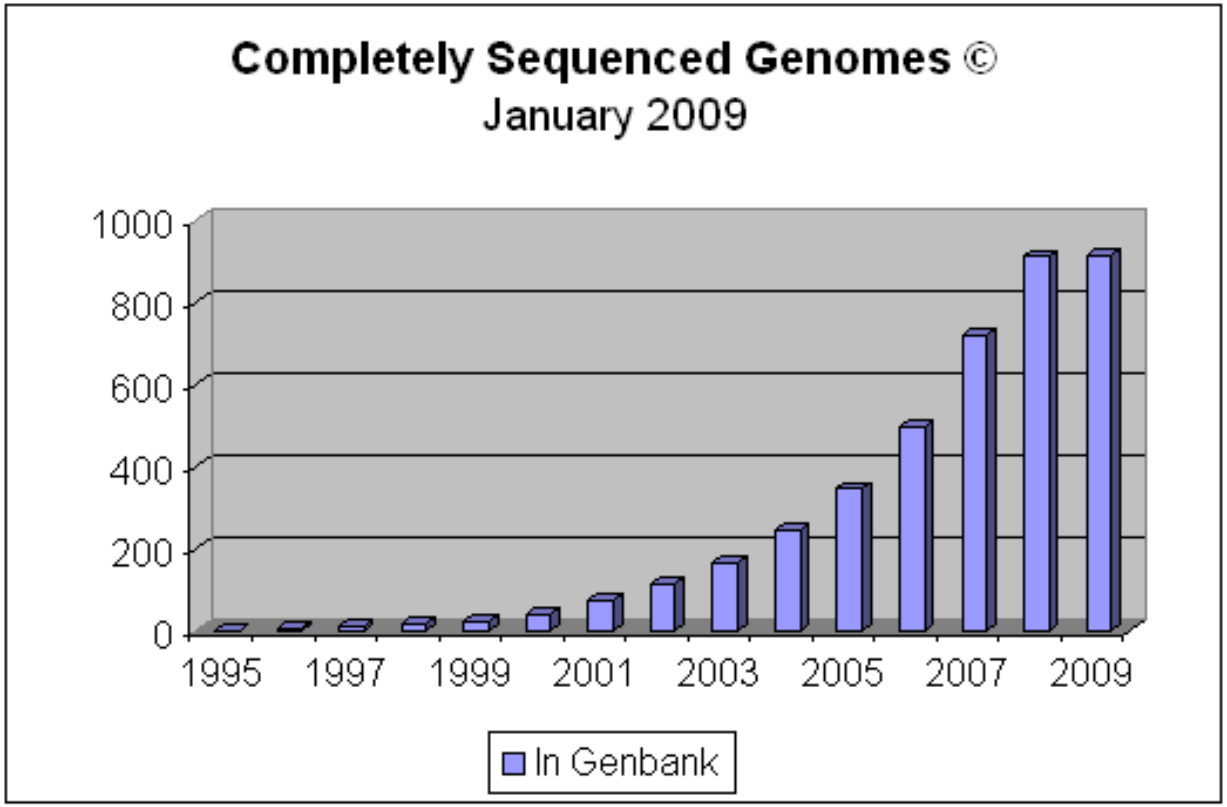
- Introdução
  - Projetos Genoma
  - Genômica Comparativa

- Métodos de Genômica Comparativa
  - Inparanoid
  - OrthoMCL
  - 3GC
  - n3GC

Introdução

- Projetos de sequenciamento de genomas
  - Sequenciamento de genoma completo de organismo permite conhecer ordem dos nucleotídeos no cromossomo:
    - características herdadas codificadas no DNA
  - Conhecimento da sequência de DNA NÃO IMPLICA conhecimento de como a informação genética leva a características e fenótipos observados
  - Foco importante de pesquisa dos projetos de sequenciamento: encontrar partes funcionais das sequências genômicas e usar estas informações para:
    - melhorar a saúde humana (individual e socialmente),
    - apoiar desenvolvimento de agricultura e pecuária, ...

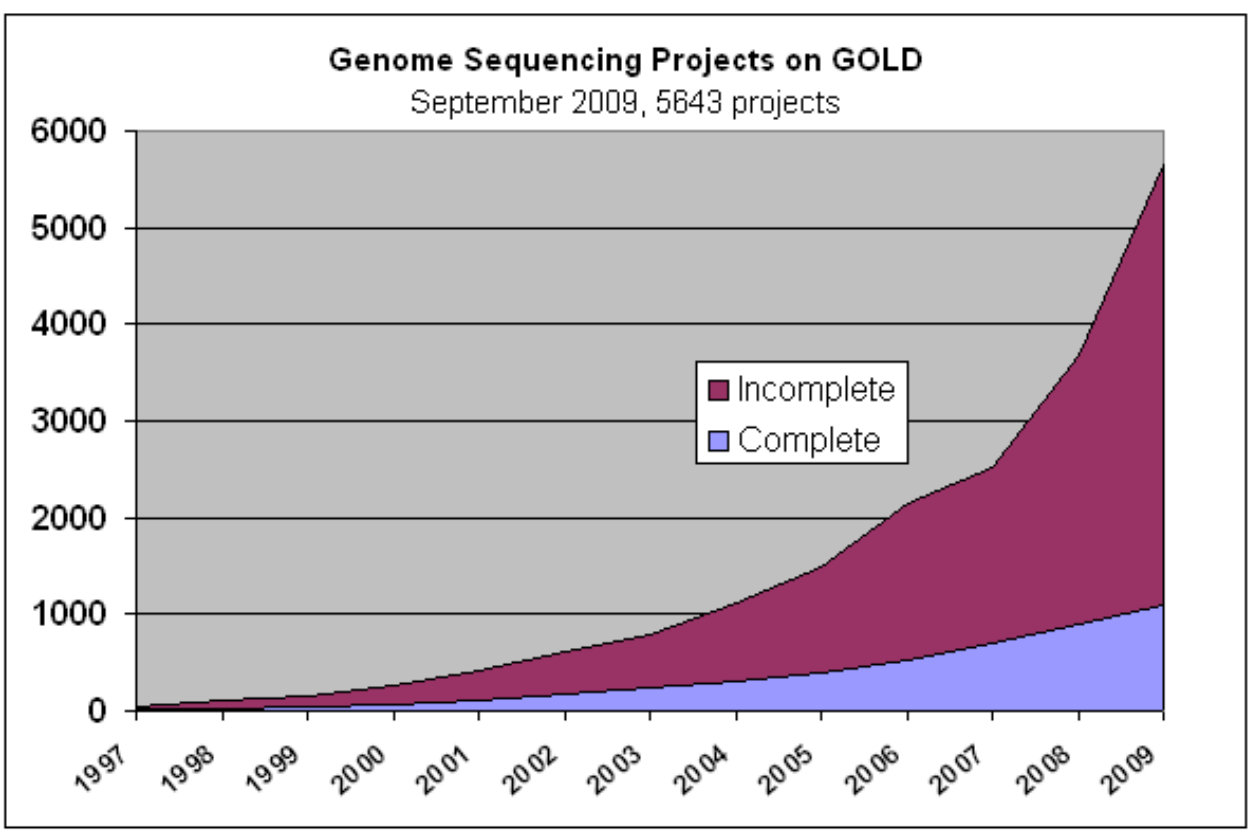
Introdução: Projetos Genoma



GOLD: Genomes Online Databases (www.genomesonline.org)

Liolios K, Chen IM, Mavromatis K, Tavernarakis N, Hugenholtz P, Markowitz VM, Kyrpides NC. The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. NAR Epub, Nov 13, 2009.

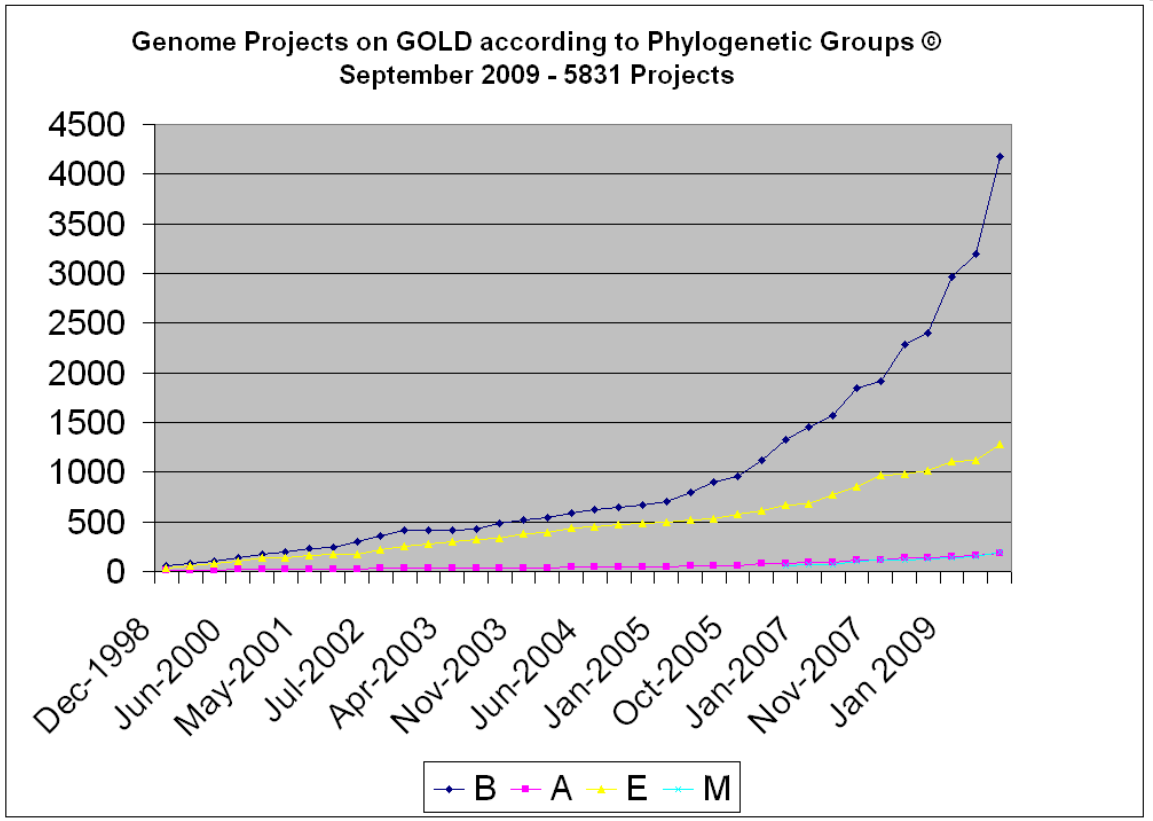
Introdução: Projetos Genoma



GOLD: Genomes Online Databases (www.genomesonline.org)

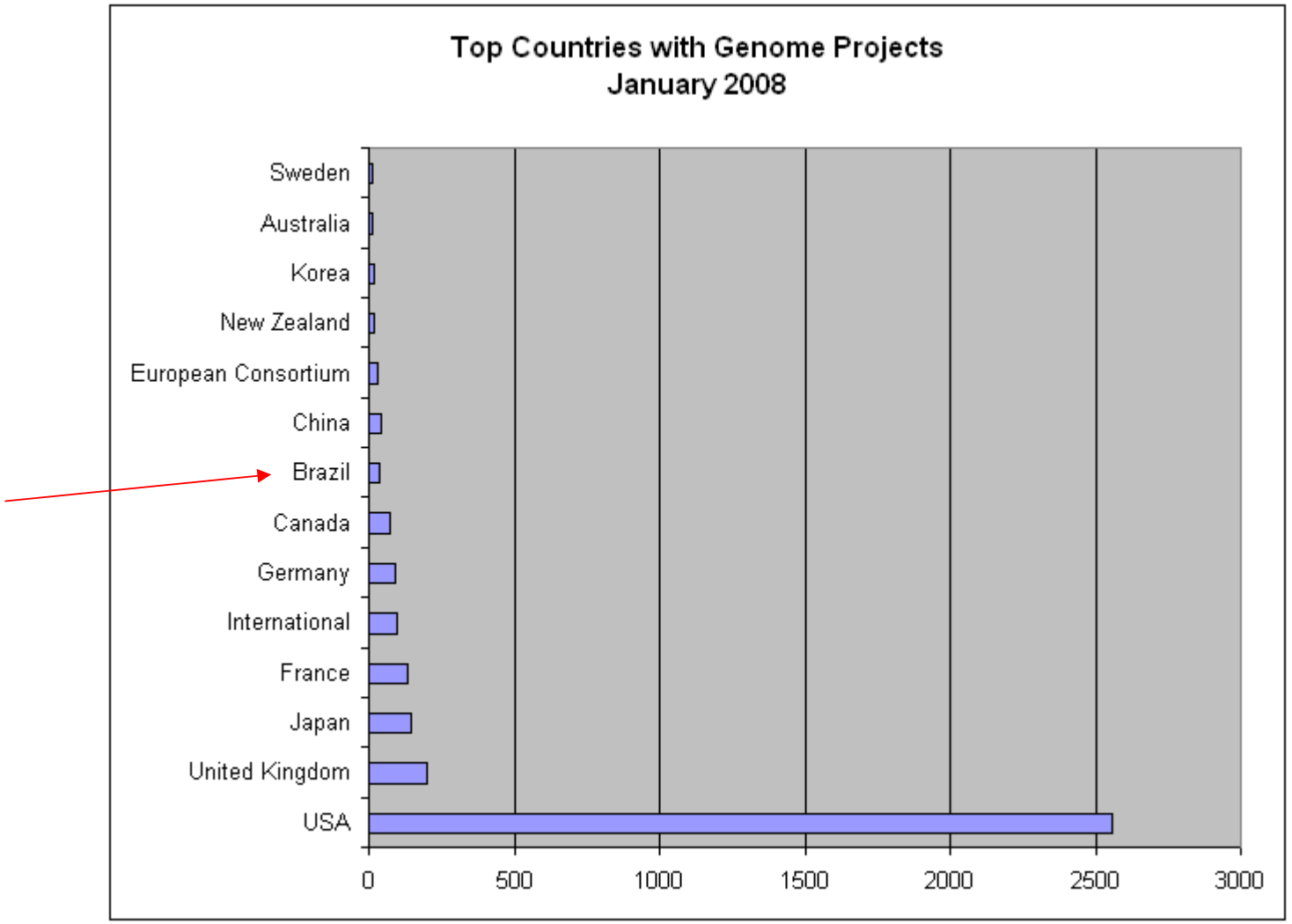
Introdução: Projetos Genoma

- Projetos genoma de acordo com grupos filogenéticos



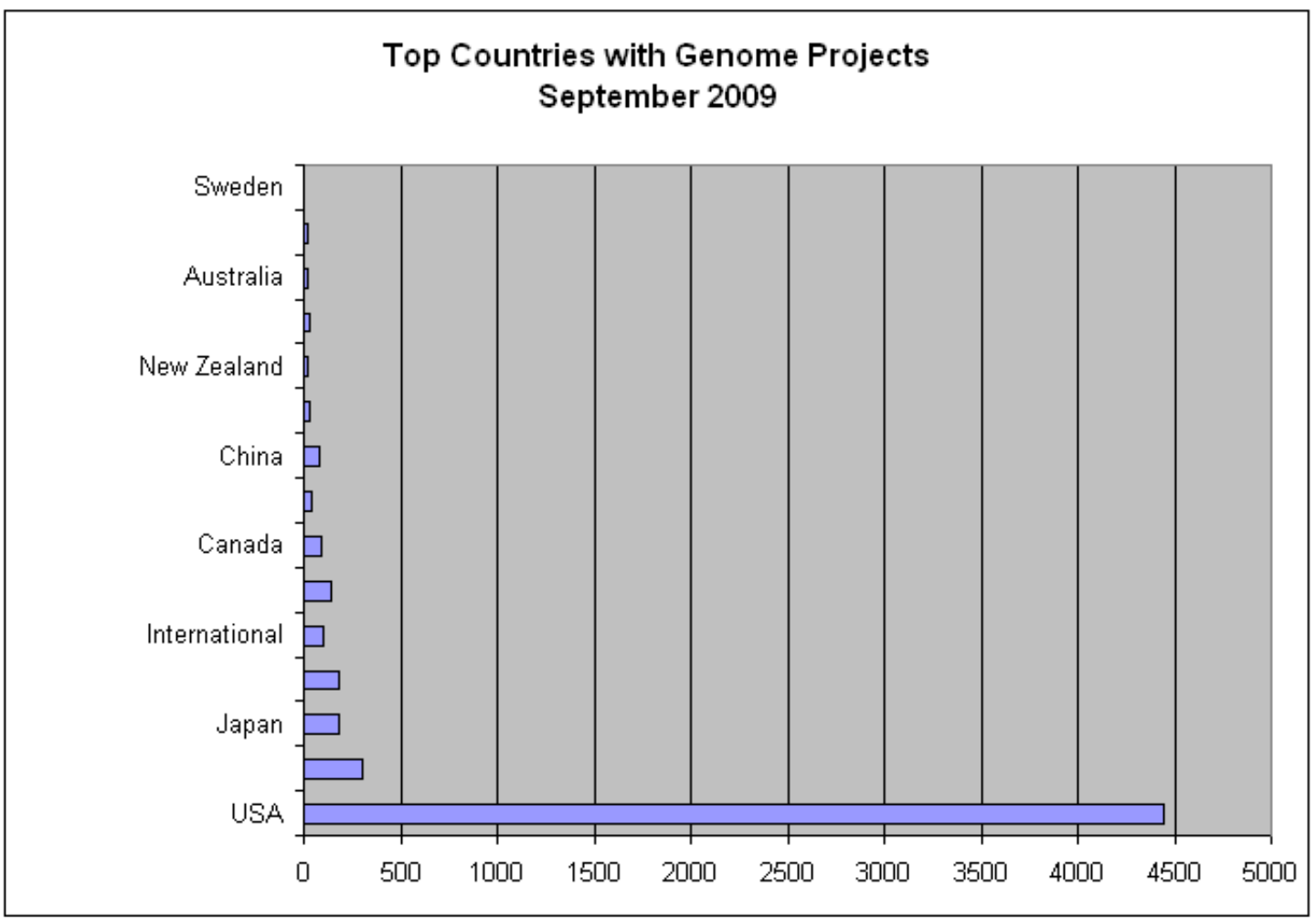
GOLD: Genomes Online Databases (www.genomesonline.org)

Introdução: Projetos Genoma



GOLD: Genomes Online Databases (www.genomesonline.org)

Introdução: Projetos Genoma



GOLD: Genomes Online Databases (www.genomesonline.org)

Introdução

- *Browsers* genômicos
  - NCBI *Map Viewer*
    http://www.ncbi.nlm.nih.gov/mapview
  - ENSEMBL
    http://www.ensembl.org
  - Universidade da California Santa Cruz (UCSC) *Genome Browser*
    http://genome.cse.ucsc.edu

Introdução

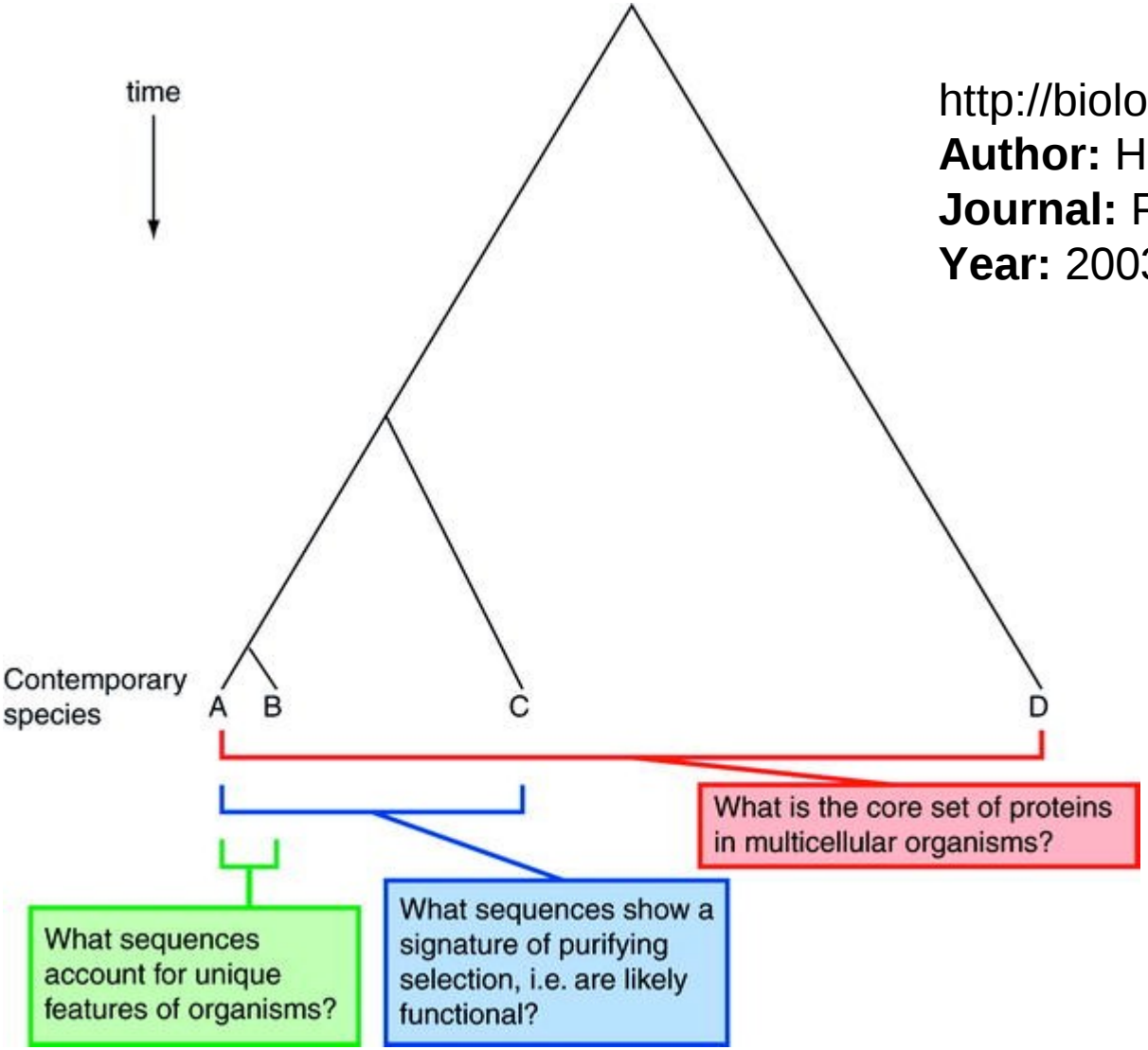
- Análises comparativas de sequências de genomas: parte essencial destas pesquisas
- Princípios de genômica comparativa:
  - características comuns entre dois organismos frequentemente codificadas no DNA, em porções conservadas entre espécies:
    - sequências de DNA codificando proteínas e RNAs responsáveis por certas funções devem ter sido:
      - conservadas a partir do último ancestral comum
      - preservadas nas espécies atuais
    - sequências de DNA regulando, de forma similar, expressão de genes entre duas espécies relacionadas devem ter sido conservadas
    - por outro lado, sequências que codificam ou controlam expressão de proteínas e RNAs responsáveis por diferenças entre espécies devem ter divergido ao longo da evolução

Introdução

- Perguntas diferentes podem ser respondidas por comparação de genomas

- Exemplo: investigar motivos para diferentes distâncias filogenéticas

Introdução



http://biology.plosjournals.org
Author: Hardison Ross C
Journal: PLoS Biology
Year: 2003 **Vol:** 1 **Issue:** 2

Introdução

Distâncias filogenéticas próximas:

- Genomas muito similares: humanos e chimpanzés (separados por cerca de 5 milhões de anos de evolução)
- Particularmente adequados para encontrar diferenças chave que podem indicar diferenças entre os organismos: mutações de sequências ocorridas por seleção positiva
- Genômica comparativa: área poderosa e muito útil à medida em que mais dados de sequências genômicas forem sendo encontradas

Introdução

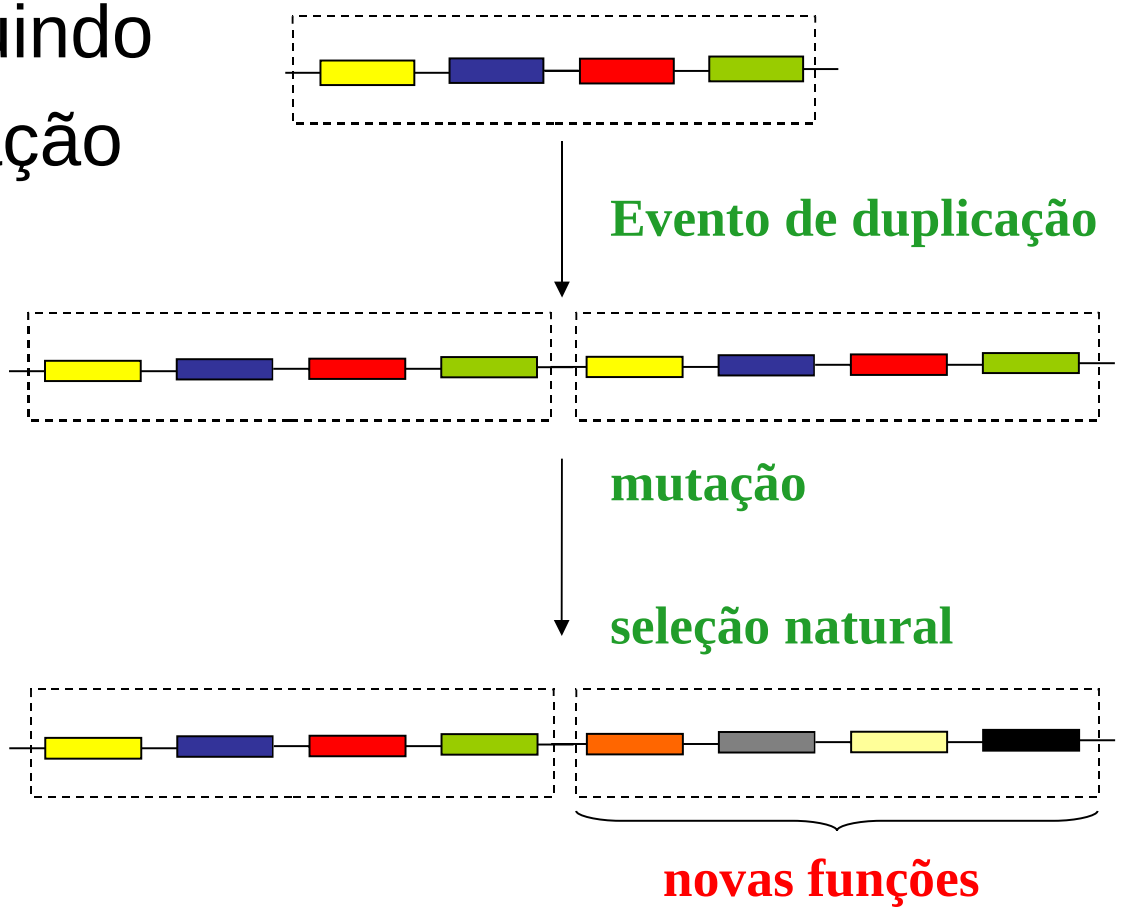
- Alinhamento de sequências de DNA é fundamental em genômica comparativa
- Alinhamento:
  - mapeamento dos nucleotídeos de uma sequência com relação aos nucleotídeos da outra sequência,
  - espaços (*gaps*) introduzidos em uma ou outra sequência para aumentar o número de posições com casamento de nucleotídeos
- Diversos algoritmos de alinhamento foram desenvolvidos para alinhar duas ou mais sequências

Introdução

- O que podemos aprender sobre a evolução?
  - Observação básica: genômica comparativa descreve casamentos entre genomas
- O que podemos aprender sobre funções num genoma?
  - Informação de similaridade de sequências entre genomas: recurso importante para encontrar regiões funcionais e para predizer estas funções
  - Exemplo: aprimoramentos na identificação de genes codificadores de proteínas
- Biólogos moleculares utilizam resultados das análises comparativas, hoje produzidas massivamente
- Aspectos foram explorados pelo Marcelo Brígido

Métodos para identificar genes parálogos e genes ortólogos

- Genes parálogos e ortólogos:
  - Genes evoluindo por duplicação

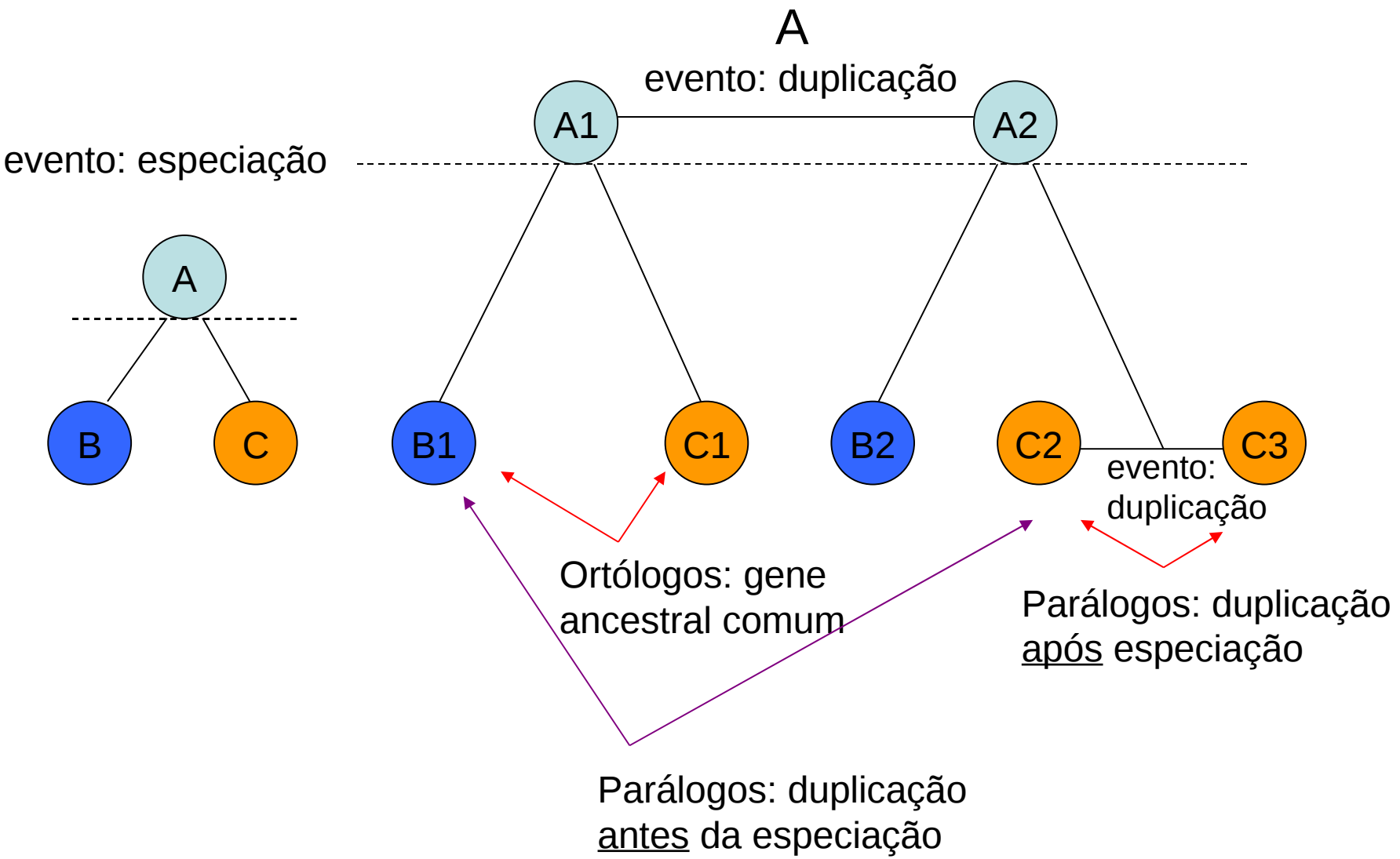


Genes parálogos e ortólogos

- Genes parálogos: originados por duplicação antes ou depois da especiação (podem possuir ou não o mesmo papel biológico)

- Genes ortólogos: originados de um único gene do último ancestral comum entre as espécies (frequentemente possuem o mesmo papel biológico dos ancestrais)

Genes parálogos e ortólogos



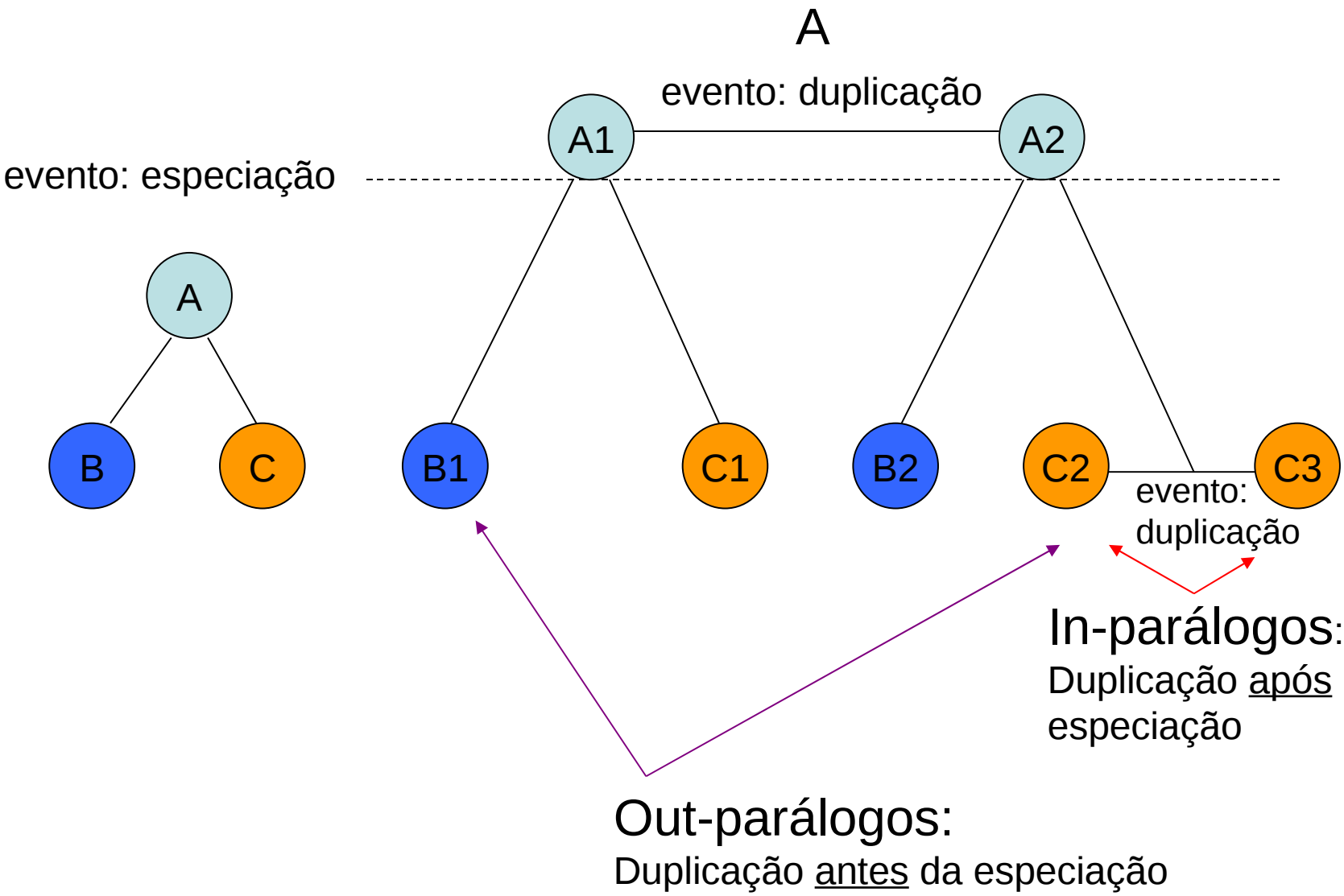
Métodos para identificar genes parálogos e genes ortólogos

- INPARANOID:
  - Remm M., Storm, C.E. V. and Sonnhammer, E. L. L. (2001). Automatic Clustering of Orthologs and In-paralogs from Pairwise Species Comparisons. Journal of Molecular Biology, 314:1041-1052.
  - Center for Genomics and Bioinformatics – Estocolmo/Suécia
  - Página: http://inparanoid.sbc.su.se

INPARANOID

- programa desenvolvido para identificar grupos de genes ortólogos verdadeiros, evitando incluir proteínas muito relacionadas, mas não ortólogas, em dois organismos:
  - in-parálogos: parálogos gerados por duplicação após especiação
  - out-parálogos: parálogos gerados por duplicação antes da especiação
  - ortólogos: não devem ser confundidos com out-parálogos

INPARANOID: Genes parálogos e ortólogos

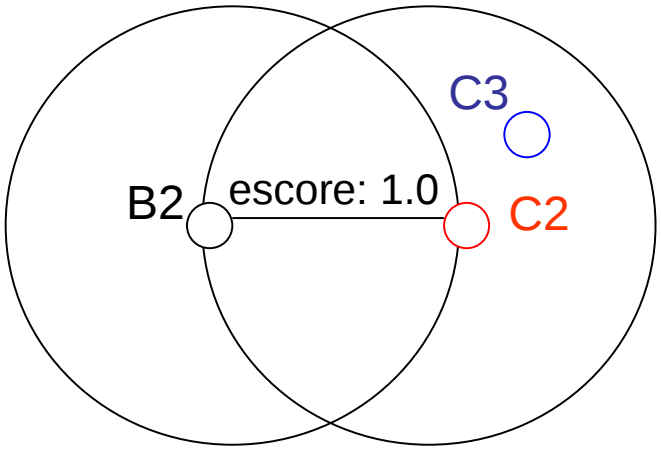


INPARANOID

- Método de clusterização:
  - Comparações BLAST entre todos os organismos dois a dois
  - Entre dois organismos A e B:
    - Organismo A com organismo B
    - Organismo B com organismo A
    - Organismo A com organismo A
    - Organismo B com organismo B
  - Grupo ortólogo:
    - Início: dois grupos constituídos por pares de ortólogos identificados por BBH (*Bi-directional Best Hit*) entre dois genomas: par de ortólogos semente
    - Sequências adicionadas a cada um dos dois grupos: sequências de cada genoma mais próximas de cada um dos ortólogos semente do que a qualquer outra sequência no outro genoma
    - Sequências incluídas num grupo de ortólogos: in-parálogos
    - Valor de confiança (escore) associado a cada sequência: mostra o seu grau de relacionamento com o ortólogo semente

INPARANOID

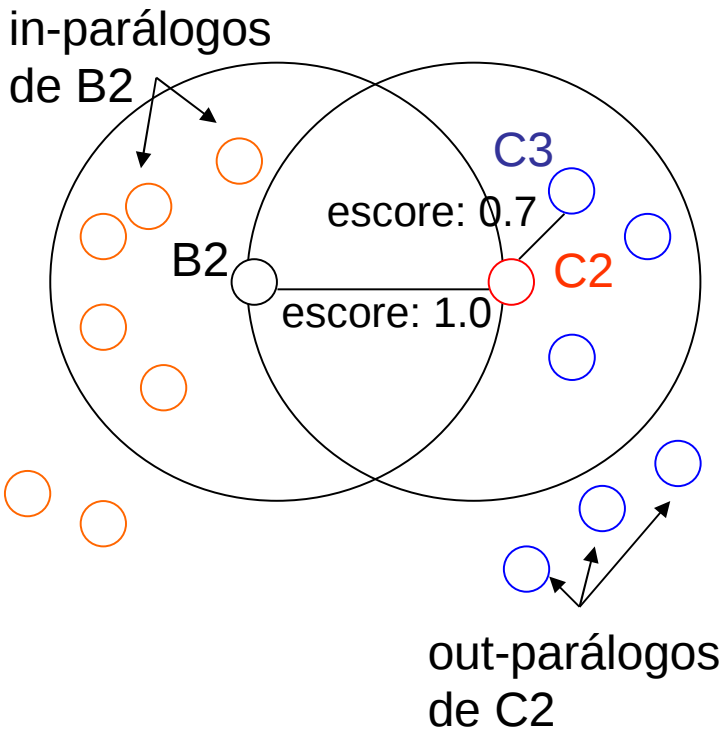
- Método de clusterização:
  - Exemplo:



B2, C2: par de ortólogos semente,
cada um recebe escore 1.0 (maior escore = menor distância),
todos os in-parálogos agrupados em torno do par (B2, C2):
C3 é in-parálogo de C2

INPARANOID

- Método de clusterização:
  - Exemplo:



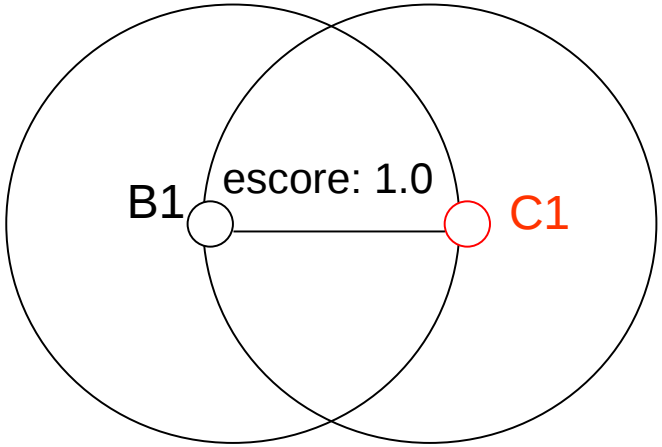
C3: escore indica similaridade relativa ao in-parálogo semente (C2):
Escore (C3) = (Blast[C2:C3] - Blast[C2:B2]) / (Blast[C2:C2] - Blast[C2:B2])
Blast[X:Y]: média do escore Blast entre X e Y (bits)

Exemplo: C2 é mais similar a B2 do que a C3, então C3 recebe escore in-parálogo menor em relação a C2: 0.7

Hipótese para agrupar in-parálogos: ortólogo principal (C2) mais similar aos in-parálogos da mesma espécie do que a qualquer outra sequência da outra espécie

INPARANOID

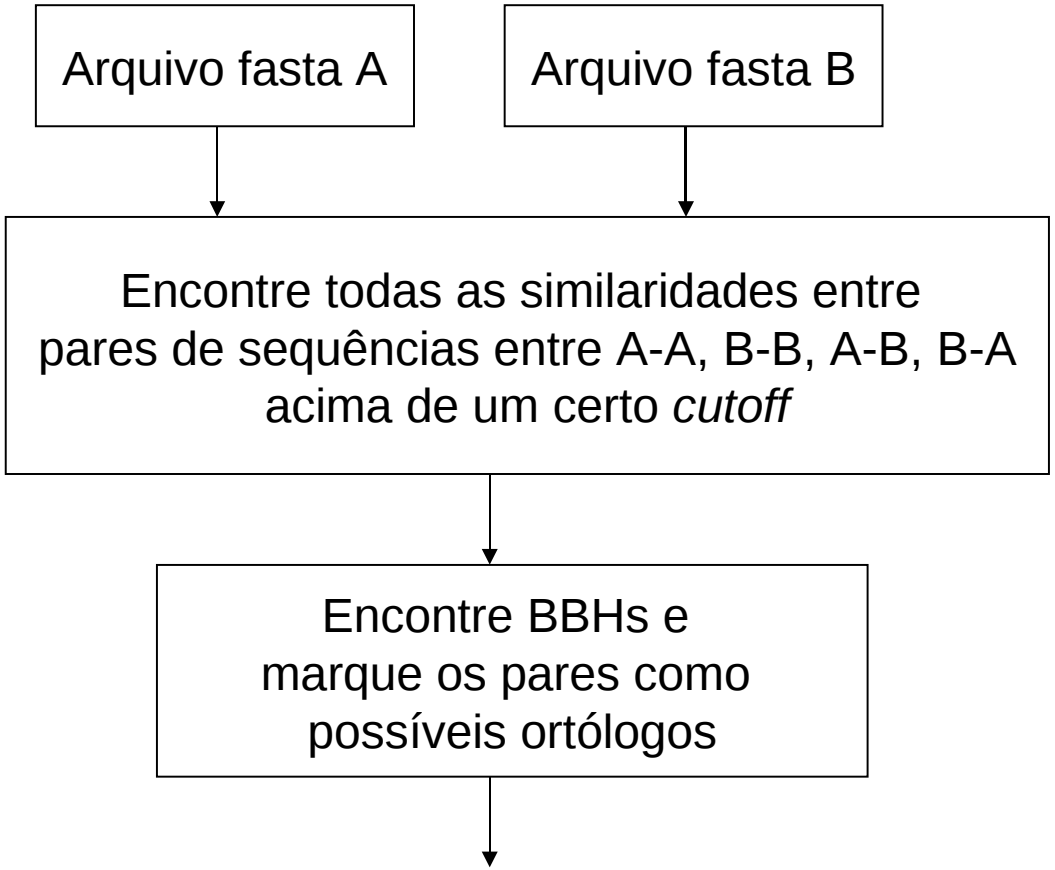
- Outro exemplo:



- B1 e C1:
  - ortólogos entre si
  - out-parálogos do grupo anterior
  - formam um grupo separado de ortólogos

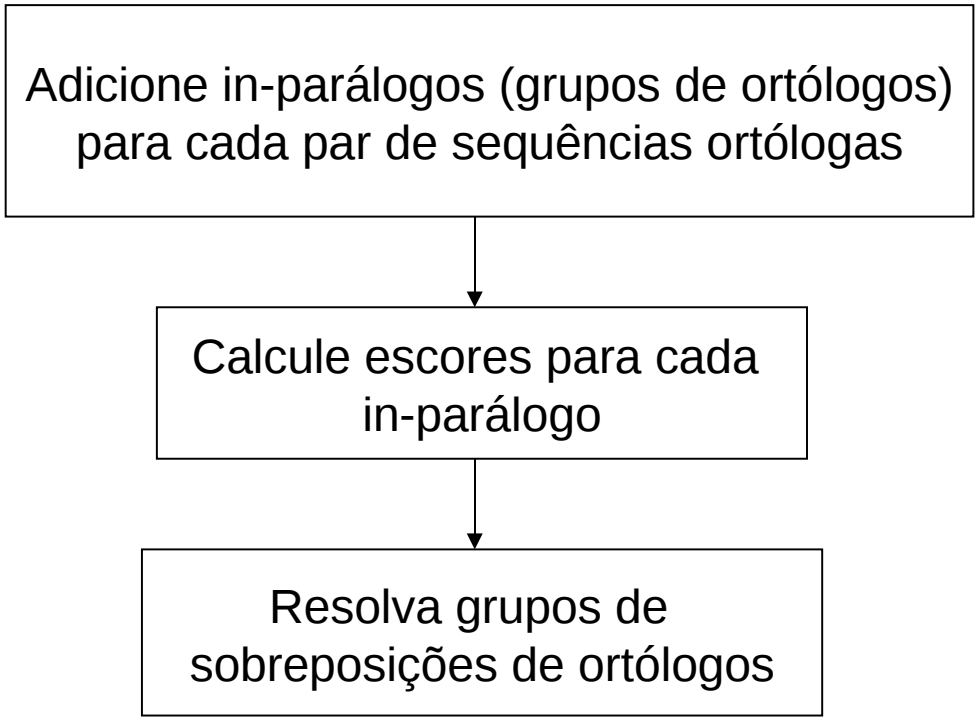
INPARANOID

- Visão geral do algoritmo (simplificado)



INPARANOID

- Visão geral do algoritmo (simplificado)



Métodos para identificar genes parálogos e genes ortólogos

- ORTHOMCL:
  - programa desenvolvido para identificar grupos de genes ortólogos em organismos eucariotos, em dois ou mais organismos
  - Christian, L. L., Stoeckert, J. and Roos, S. (2003). OrthoMCL: Identification of Ortholog Groups for Eukariotic Genomes. Genome Research, 13(9):2178-2189.
  - Center for Genomics and Bioinformatics – Estocolmo/Suécia
  - Página: http://www.orthomcl.org

OrthoMCL

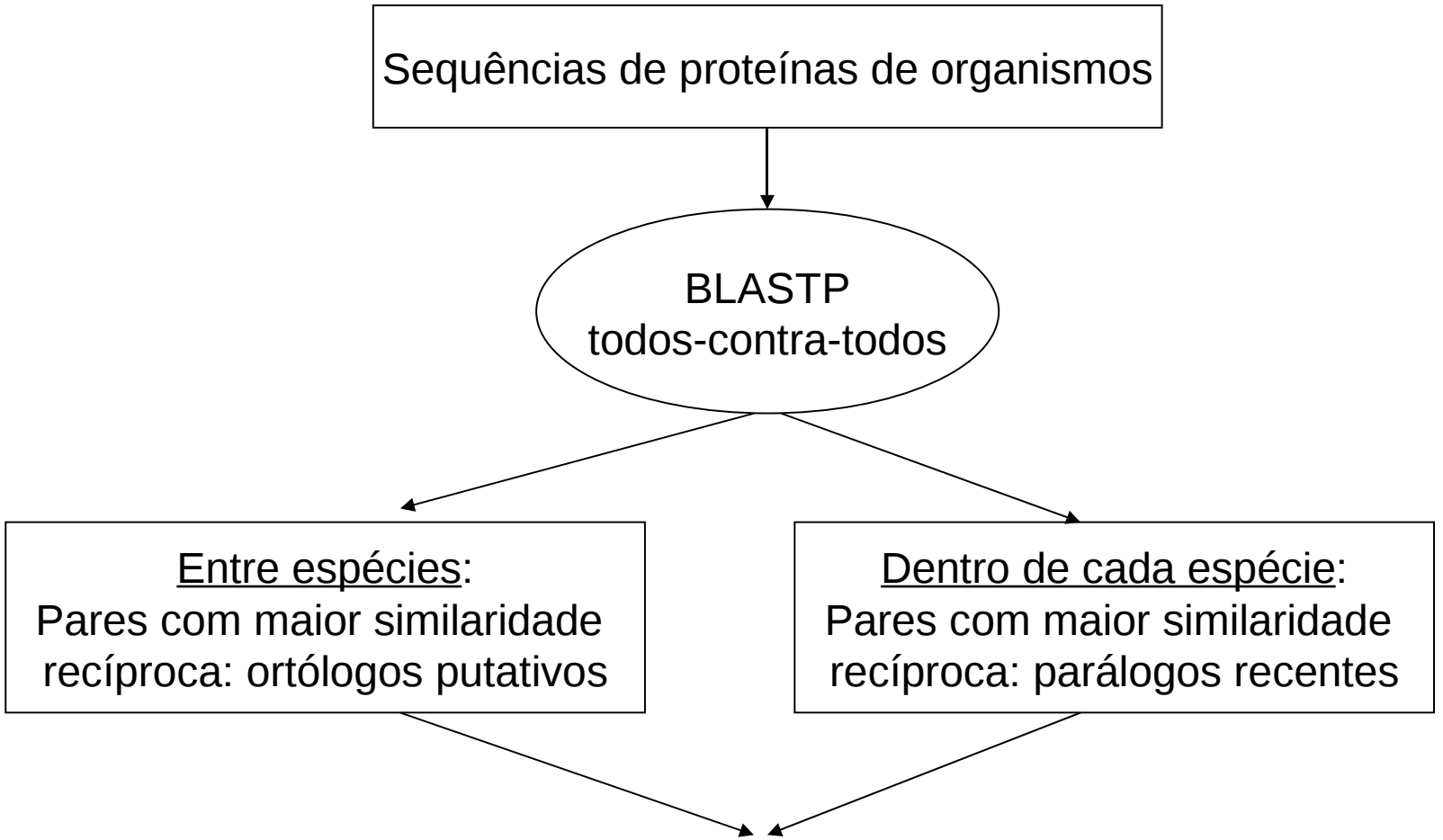
- Distingue redundância funcional (duplicação) e divergência:
  - identifica parálogos recentes (in-parálogos do INPARANOID) a serem incluídos em grupo de ortólogos por *hits* Blast melhores na mesma espécie quando comparados a *hits* entre espécies diferentes
- Similar ao INPARANOID, difere: parálogos recentes devem ser mais similares entre eles do que a qualquer sequência de outra espécie
- Resolver relacionamentos ortólogos muitos-para-muitos surgidos de comparações entre múltiplos genomas: algoritmo de Markov Cluster (MCL)
- MCL: simula caminhos aleatórios num grafo utilizando matrizes de Markov para determinar probabilidades de transição entre os vértices do grafo

OrthoMCL

- Gera grupos de proteínas
- Grupo: ortólogos ou parálogos recentes de pelo menos 2 espécies

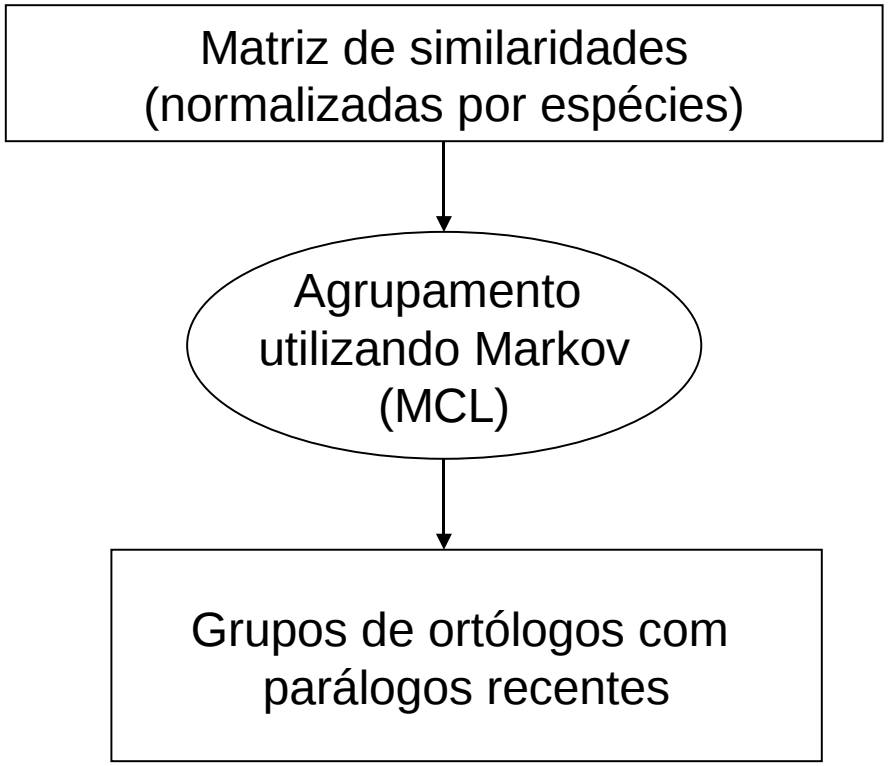
OrthoMCL

- Método para agrupar proteínas ortólogas



OrthoMCL

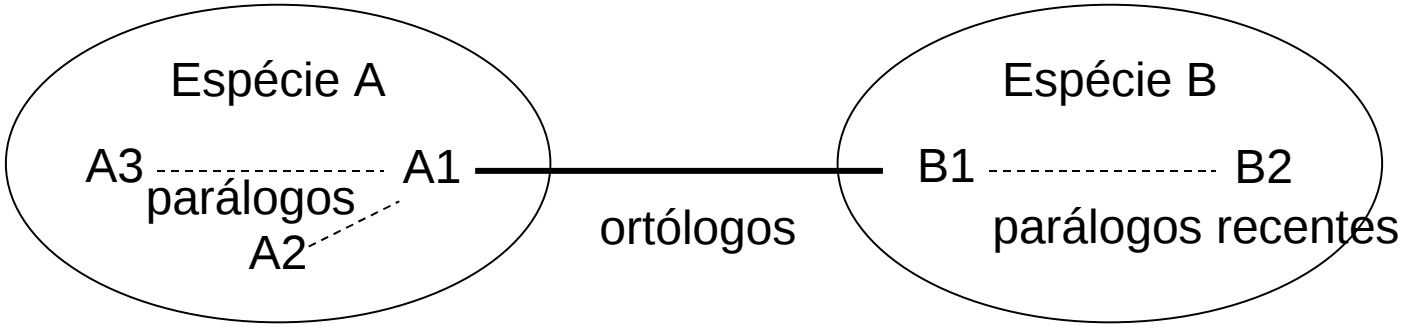
- Método para agrupar proteínas ortólogas



OrthoMCL

- Ortólogos putativos e parálogos recentes representados em um grafo, onde:
  - Vértices: representam sequências de proteínas
  - Arestas ponderadas: relacionamentos entre as proteínas
- Peso = - $\log_{10}$ (*P-value*)
  - *P-value*: resultado Blast para cada par de sequências
- Alta similaridade de parálogos recentes pode alterar processo de agrupamento: pesos das arestas normalizados para refletir o peso médio para todos os pares de ortólogos destas duas espécies (ou dentro da mesma espécie para parálogos recentes)
- Grafo resultante: matriz de similaridades não é simétrica
- MCL aplicado à matriz de similaridades

Exemplo: relacionamentos entre duas sequências e matriz de similaridades



	A1	A2	A3	B1	B2
A1		300	152	61	29
A2	233		150	60	29
A3	118	117		40	30
B1	69	68	45		88
B2	33	33	34	100	

Pesos normalizados corrigidos por MCL

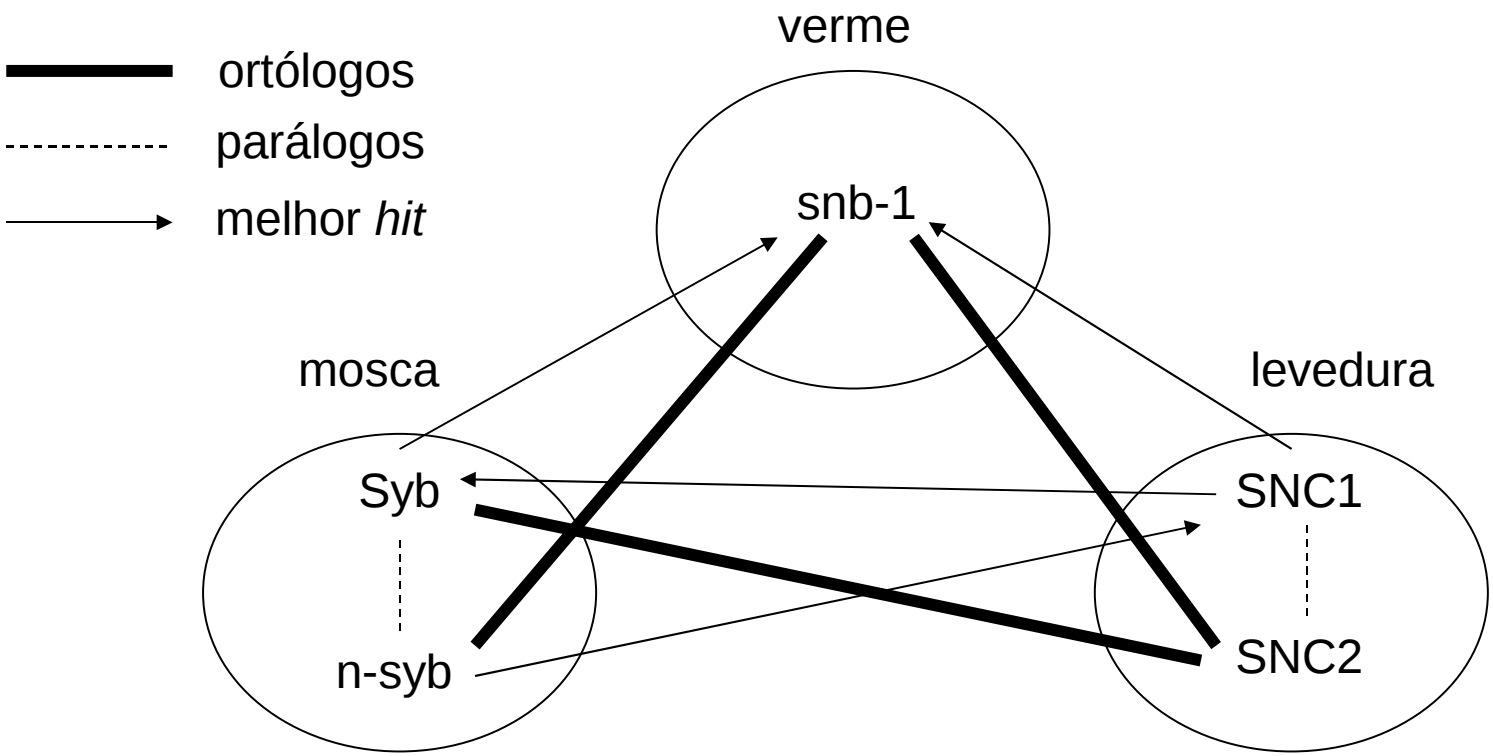
Pesos iniciais:
$-\log_{10}$ (*P-value*)
similaridades WU-BLASTP

OrthoMCL

- MCL:
  - usa simulação de fluxo
  - considera todos os relacionamentos no grafo globalmente e simultaneamente durante o agrupamento
  - método robusto para separar:
    - parálogos que divergiram
    - ortólogos distantes incorretamente identificados baseados em BBHs fracos
  - Parâmetro importante - valor *inflating*:
    - controla o tamanho do grupo (granularidade)
    - aumentar o valor: aumenta o tamanho do grupo

OrthoMCL

- Saída: grupos com sequências de pelo menos duas espécies – ortólogos e parálogos recentes



Métodos para comparar três genomas: 3GC

- Telles, G. P., Brígido, M. M., Almeida, N. F., Viana, C. J., Anjos, D. A. S., Walter, M. E. M. T. (2005). A method for comparing three genomes. Lecture Notes in Bioinformatics, Alemanha, v. 3594, p. 160-169.

- Página: http://egg.dct.ufms.br/3gc

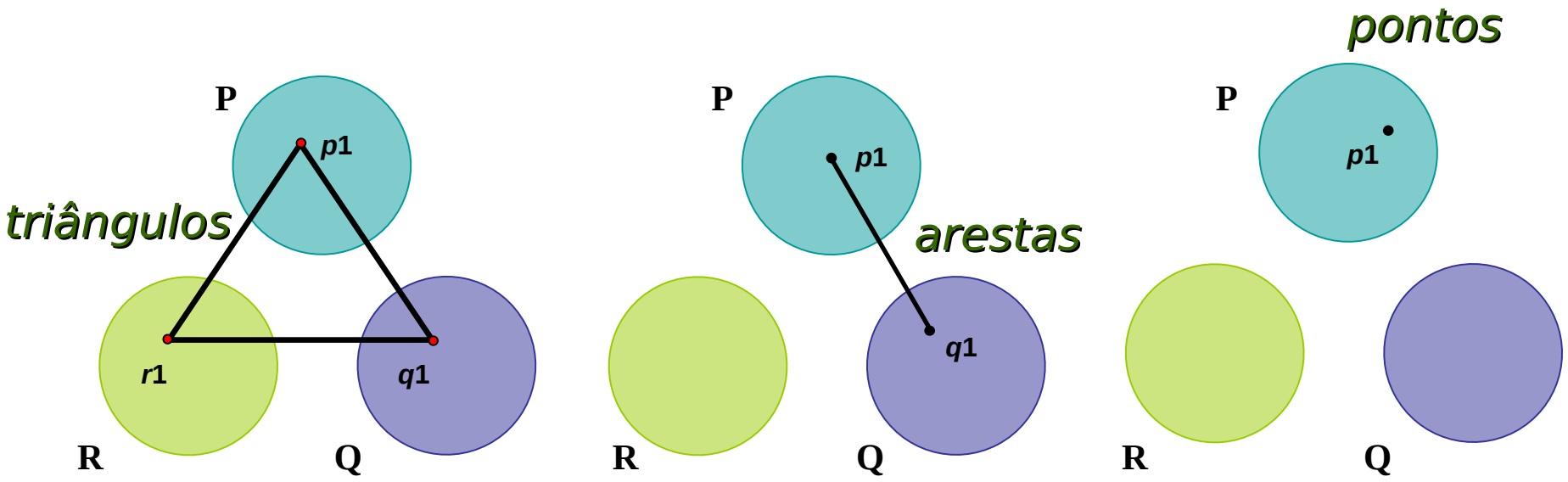
Métodos para comparar três genomas: 3GC

- Objetivo: Propor um método para comparar três genomas, simultaneamente, para descobrir características comuns entre eles

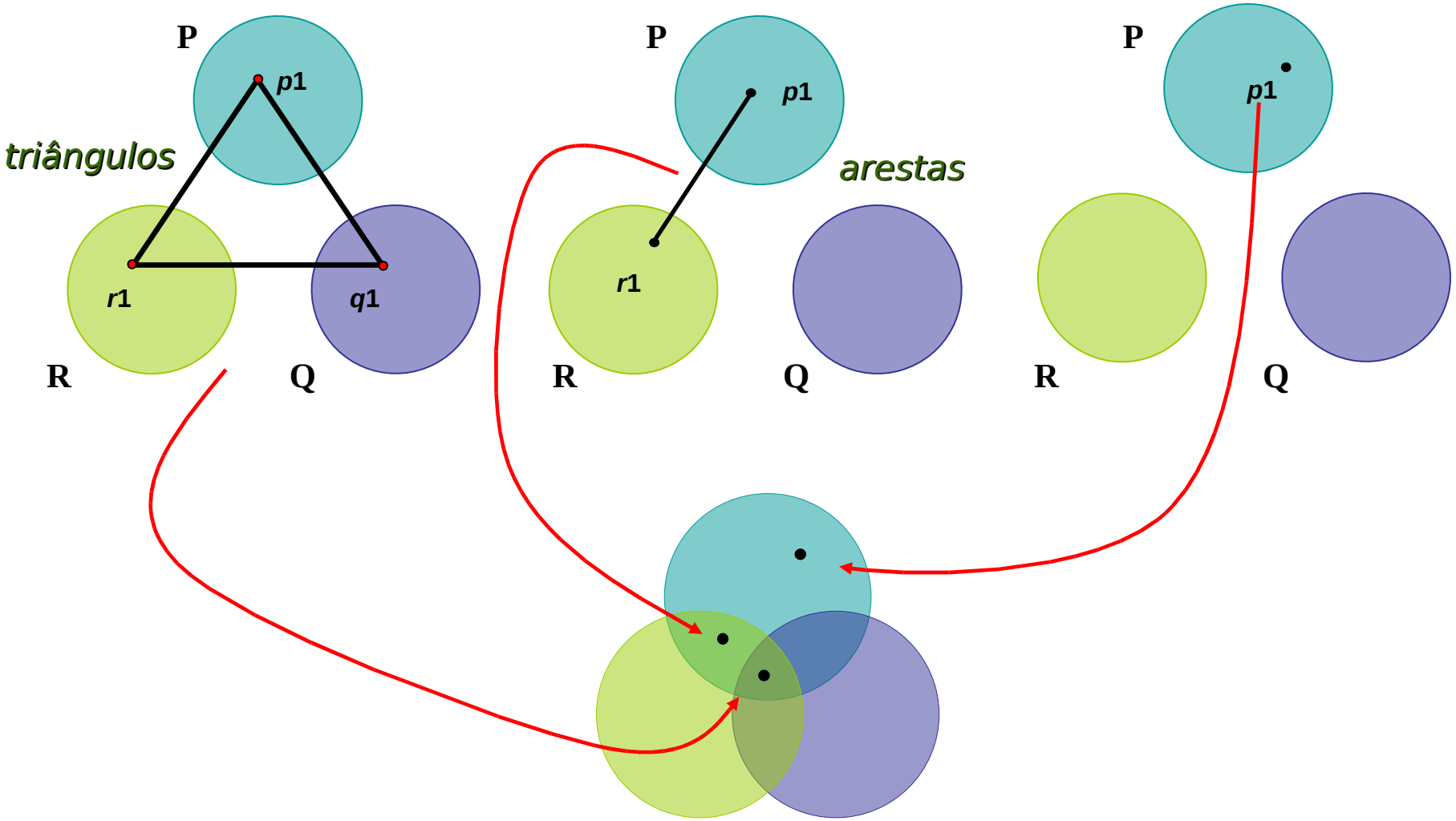
- Genoma: conjunto de sequências que podem ser:
  - DNA codificador ou
  - polipeptídeo predito

Método 3GC

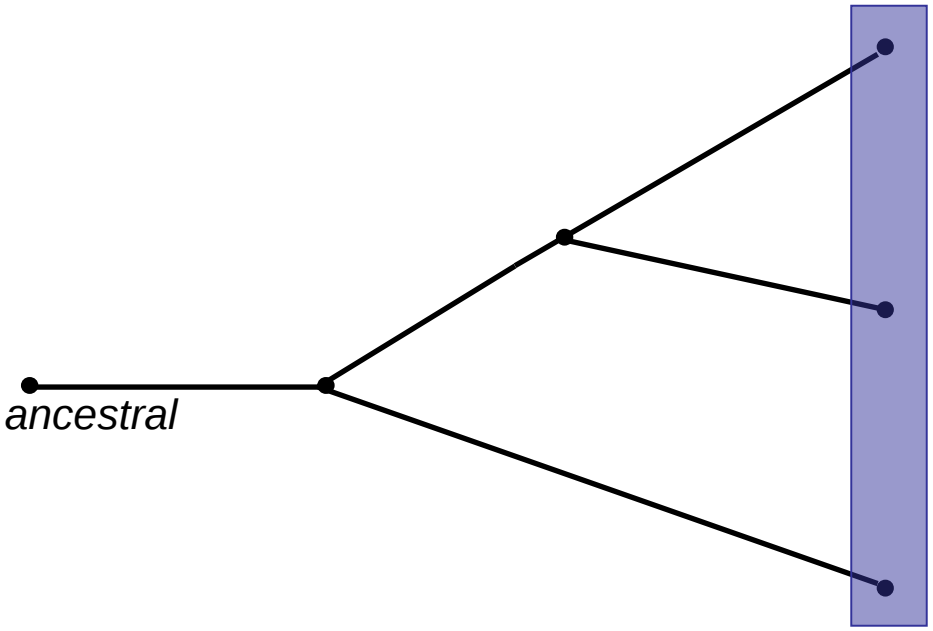
Triângulos, arestas e pontos



Triângulos, arestas e pontos



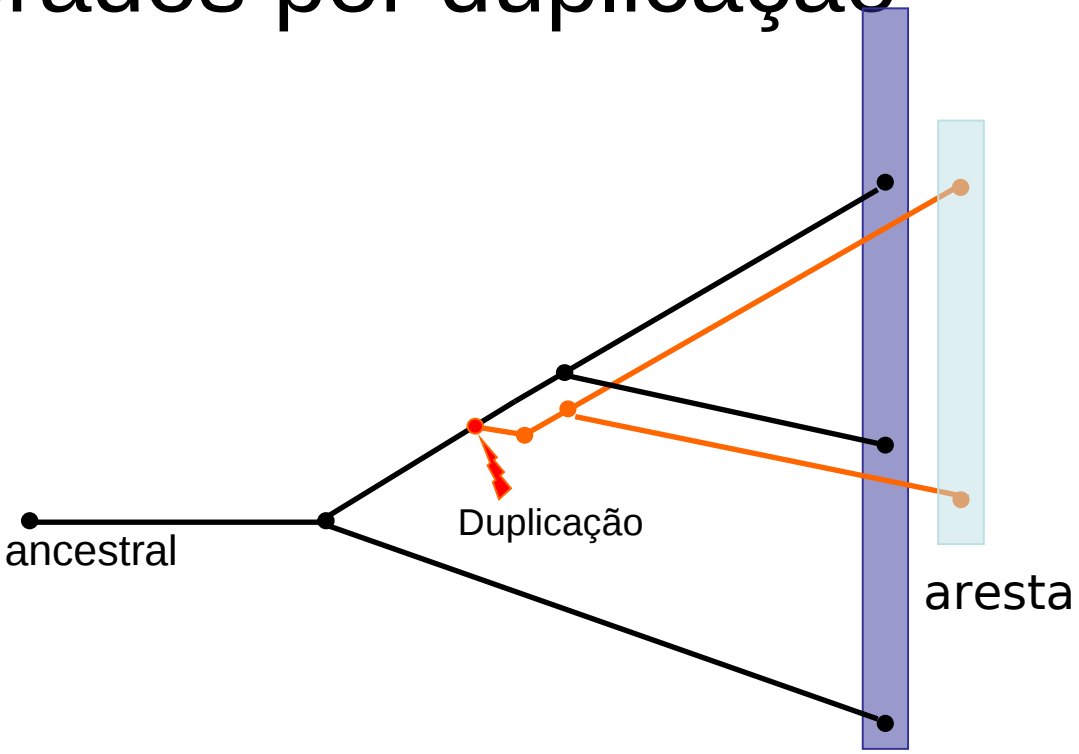
Triângulos representam grupos de ortólogos



Triângulo ortólogo

Arestas representam genes ausentes em um genoma

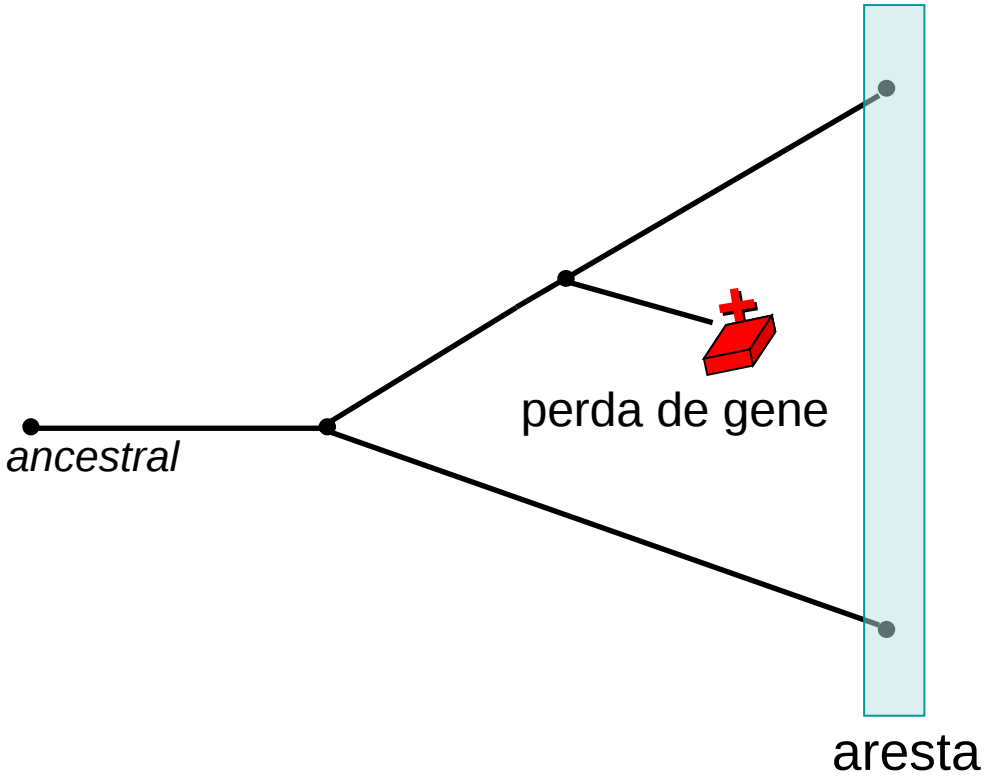
- Gerados por duplicação



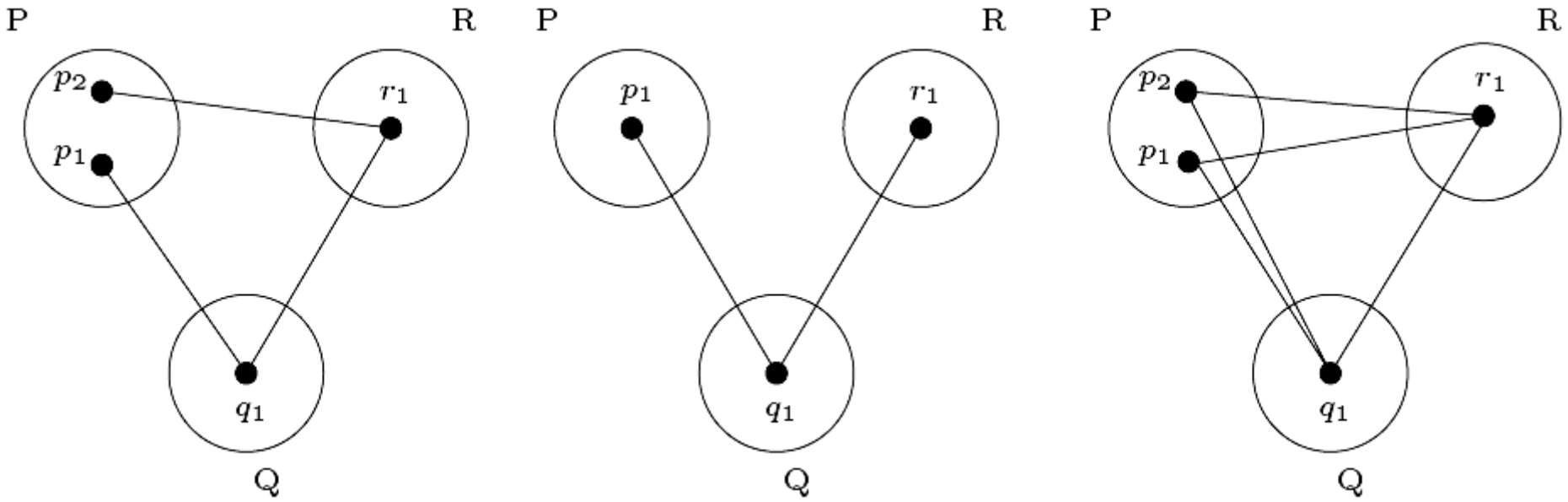
Triângulo relacionado

Arestas representam genes ausentes em um genoma

- ou devido à perda de gene

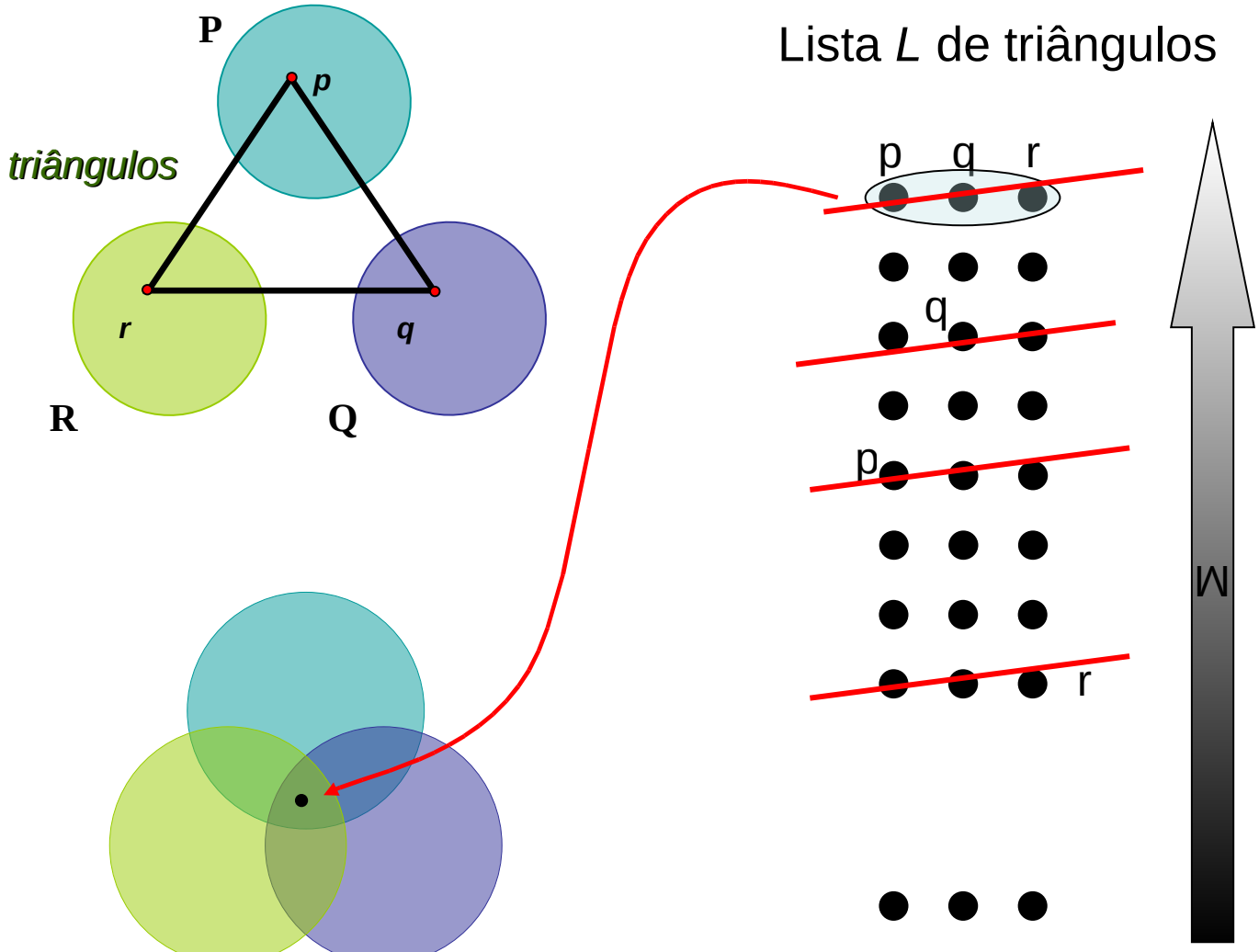


Relacionamentos ambíguos

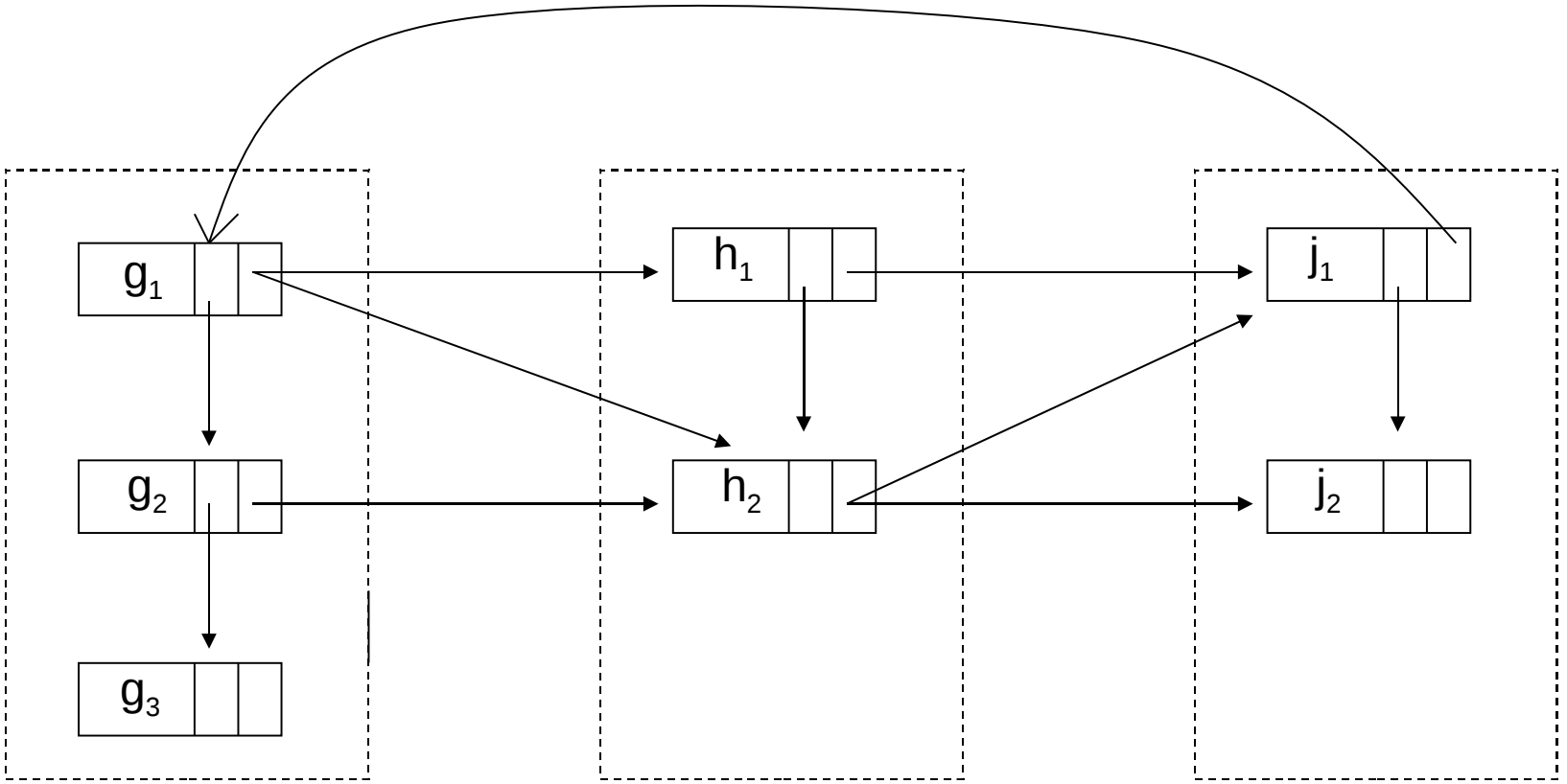


Outros relacionamentos complexos podem aparecer

Método



Método 3GC: Estruturas de Dados



triângulo: {g_1, h_1, j_1} ou {g_1, h_2, j_1}

aresta: {g_2, h_2}

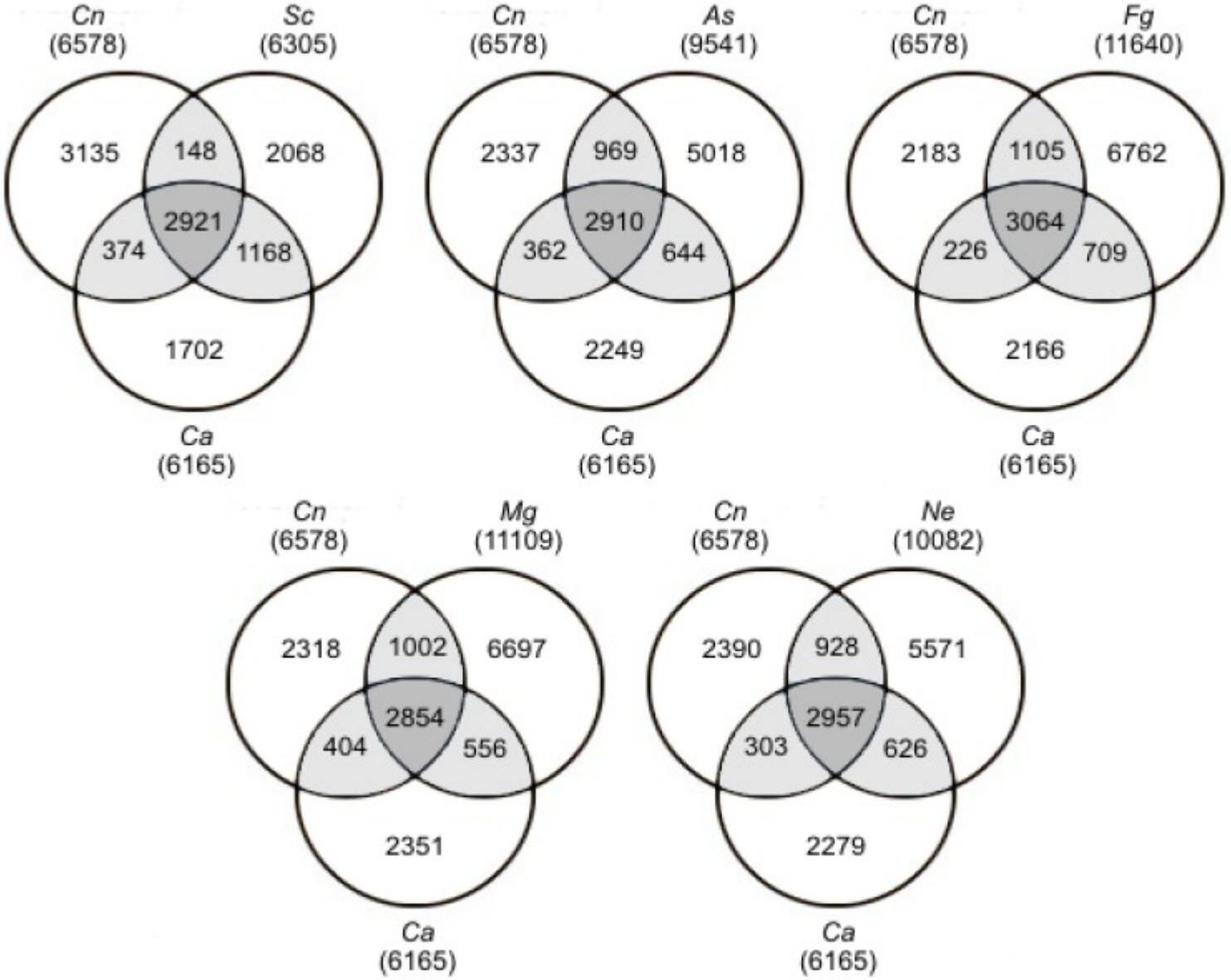
nós: {g_3} e { j_2}

Complexidade de tempo e espaço

- Tempo: $O(g^3 m^2)$
- Espaço: $O(g^3)$

Experimento: comparação genômica

- *Candida albicans*
- *Criptococcus neoformans*
- *Aspergillus nidulans*
- *Saccharomyces cereviseae*
- *Neurospora crassa*
- *Magnaporthe grisea*
- *Fusarium graminearum*


Cn
(6578)
Sc
(6305)
3135
148
2068
2921
374
1168
1702
Ca
(6165)
Cn
(6578)
As
(9541)
2337
969
5018
2910
362
644
2249
Ca
(6165)
Cn
(6578)
Fg
(11640)
2183
1105
6762
3064
226
709
2166
Ca
(6165)
Cn
(6578)
Mg
(11109)
2318
1002
6697
2854
404
556
2351
Ca
(6165)
Cn
(6578)
Ne
(10082)
2390
928
5571
2957
303
626
2279
Ca
(6165)

Método 3GC

- Método geral para comparar 3 genomas
  - Entrada: 3 conjuntos de sequências de genes
  - Saída: diagrama de Venn-Euler representando genes: exclusivos, comuns a pares de genomas, e comuns aos 3 genomas
- Tempo razoável
- 90% dos triângulos: no mínimo 2 Pfams idênticos (80% com 3 Pfams idênticos)
- Confirmados por análises filogenéticas

Problemas

- 3GC agrupa genes na ordem em que estes genes são encontrados:
  - Genes homólogos separados em grupos diferentes
- Não distingue parálogos entre espécies diferentes de ortólogos

Métodos para comparar três genomas: n3GC

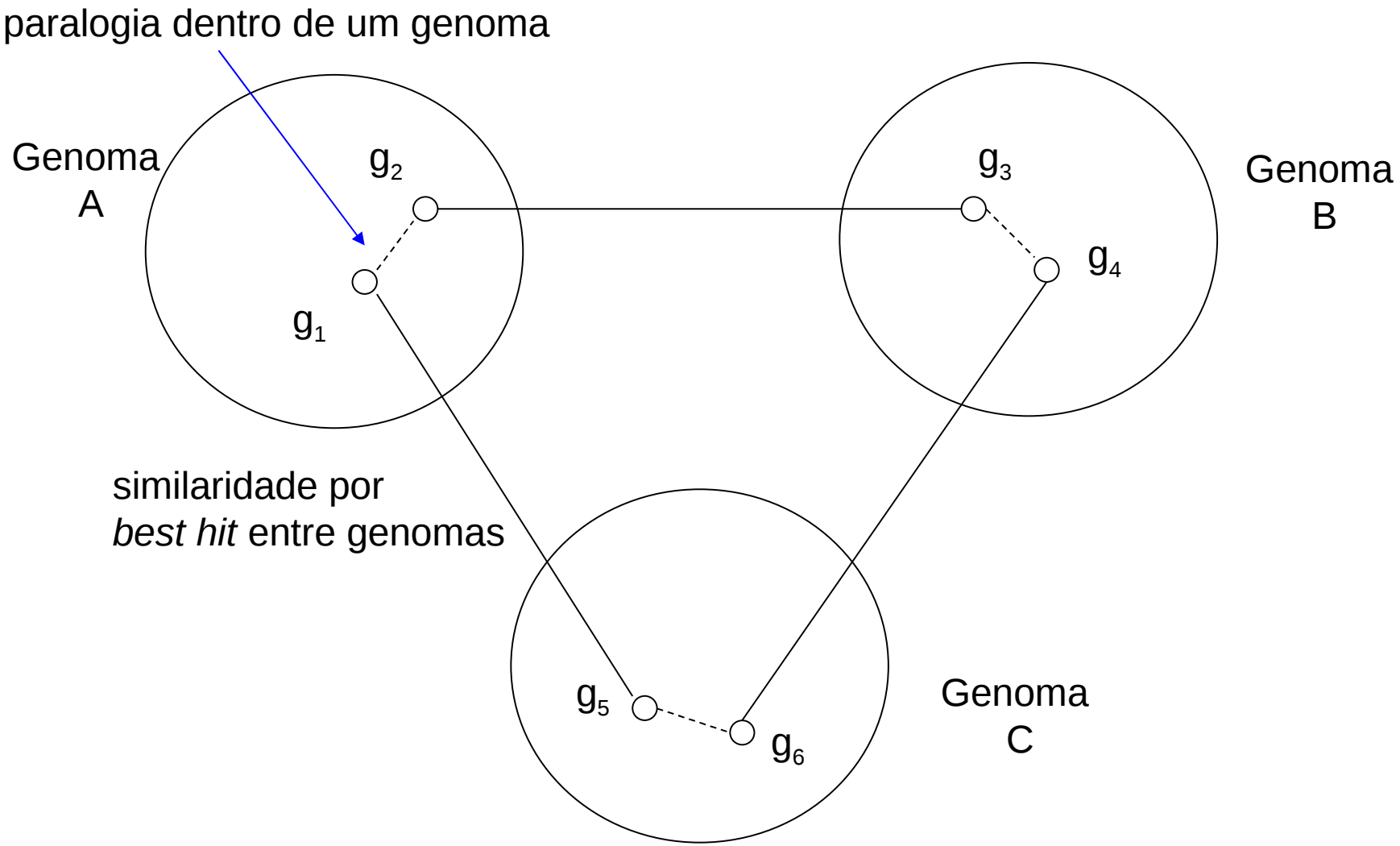
- Anjos, D. A. S., Zerlotini, G. G., Pinto, G. A., Walter, M. E. M. T., Brígido, M. M., Telles, G. P., Viana, C. J., Almeida, N. F. (2007). A method for inferring biological function using homologous genes among three genomes. In: BSB 2007, Angra dos Reis. LNBI/Advances in Bioinformatics and Computational Biology. Springer, 69-80.

- Página: http://egg.dct.ufms.br/n3gc

n3GC

- Problemas do 3GC:
  - pode colocar genes parálogos nas interseções (triângulos, arestas, nós): não identificados como parálogos
  - pode colocar pares de genes na interseção de dois genomas, que poderiam ser colocados na interseção dos três genomas

n3GC


paralogia dentro de um genoma
Genoma A
g2
g1
g3
g4
Genoma B
similaridade por
best hit entre genomas
g5
g6
Genoma C

n3GC

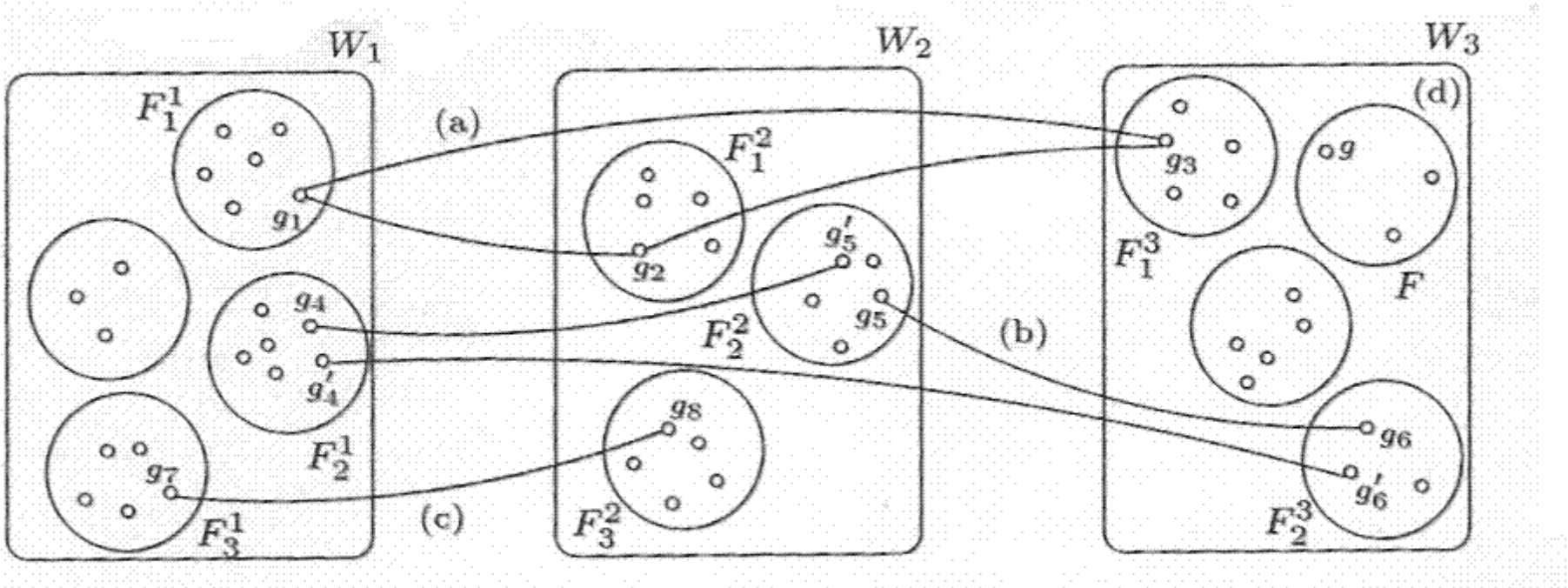
- Problema: $\{g_1, g_2, \ldots, g_6\}$ poderiam não ser reconhecidos como comuns aos três genomas
  - resultado esperado se paralogia dentro de cada genoma fosse considerado
  - identificação de paralogia <u>após</u> a aplicação do método 3GC:
    - necessitaria comparações entre todos os genes de todas as interseções encontradas
  - Método n3GC:
    - paralogia identificada <u>antes</u> das comparações entre três genomas

n3GC

- Objetivo: Propor um método para comparar três genomas simultaneamente, para descobrir características comuns entre eles, mas agrupando os genes parálogos no mesmo genoma

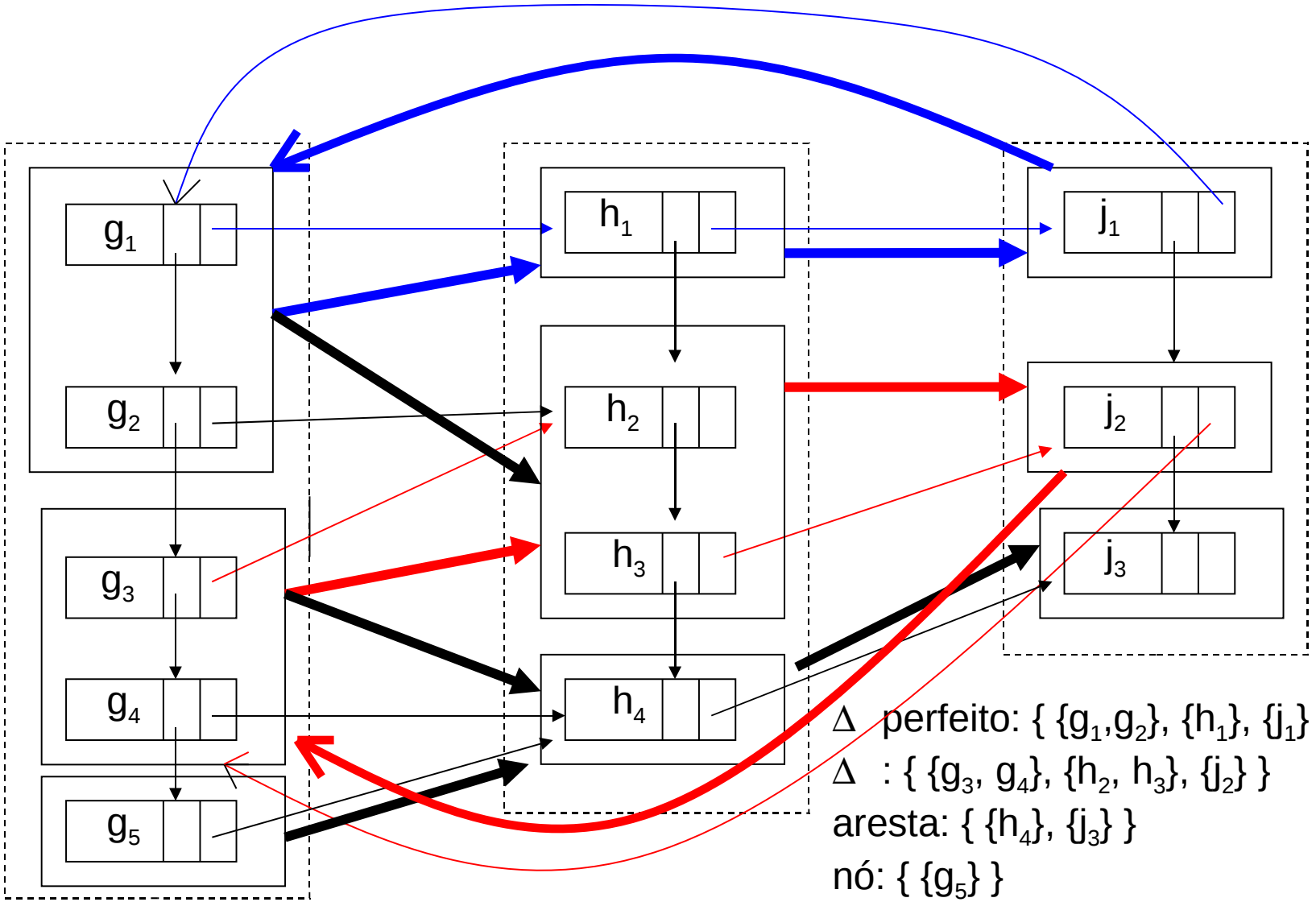
- Genoma: conjunto de sequências que podem ser DNA codificador ou polipeptídeos preditos

n3GC: Estruturas de dados



Grafo tripartido: cada partição marcada por um quadrado
Famílias: mostradas dentro dos círculos
(c) Triângulos perfeito
(d) Triângulos
(e) Arestas
(f) Nós

n3GC: Estruturas de dados



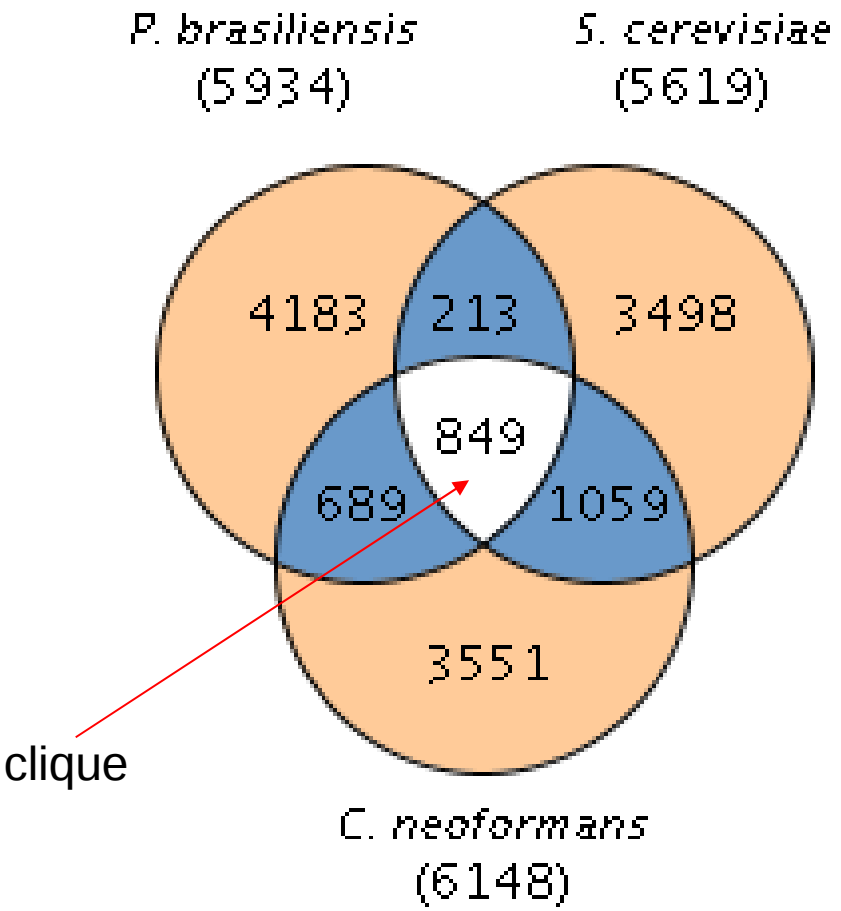
Método n3GC

Entrada: famílias de genes parálogos de 3 genomas

Saída: triângulos perfeitos, triângulos, arestas, nós

5. Encontre todos os triângulos perfeitos
6. Encontre todos os triângulos
7. Encontre todas as arestas
8. Encontre todos os nós

n3GC: Visualização dos resultados



n3GC

PBgi|51773 Contig1773 GTPases
PBgi|52267 Contig2267 involved in the secretion pathway at the ER-to-Golgi ...
PBgi|55456 PBGEX-Y1-074t_F02 secretion related GTPase
PBgi|52145 Contig2145 RAS-RELATED PROTEIN RAB-1A
PBgi|52619 Contig2619 GTP-binding protein
PBgi|52734 PBDCR-M1-011t_C06 Ras GTPase superfamily
++++++++++
CNgi|509131 cn09131 CNBI1310 hypothetical protein
++++++++++
SCgi|6320869 ref|NP_010948.1| probably involved in intra-Golgi transport ...
SCgi|6321228 ref|NP_011305.1| probably involved in intra-Golgi transport ...
SCgi|14318480 ref|NP_116615.1| involved in the secretion pathway at the ...
SCgi|14318517 ref|NP_116650.1| Involved in transport and fusion of post-Golgi ...
SCgi|6323291 ref|NP_013363.1| Ras-like GTP binding protein involved in the ...
SCgi|6323642 ref|NP_013713.1| Gtp-binding protein of the rab family; required for ...
++++++++++
PBgi|51773 -> CNgi|509131
CNgi|509131 -> SCgi|14318480
SCgi|14318480 -> PBgi|51773

n3GC: Análise dos resultados

- Comparações com PFam:
  - 95% dos relacionamentos com 1 ou 2 PFams: mesma função ou similar
- Comparações com INPARANOID e 3GC:
  - resultados similares
- Desempenho prático bom:
  - Pentium PC 4 1.5GHz, 512 MB memória: 9,5 minutos em média
  - resultados BLAST já obtidos
  - Tamanho médio dos genomas: 8.500 genes

n3GC: Trabalhos Futuros

- Identificação de ortólogos e parálogos em espécies diferentes
- Disponibilizar página para uso público
- Executar BLAST de forma distribuída
- Flexibilizar o uso de ferramentas para comparação entre genomas