TESE DE DOUTORADO

Desenvolvimento das ferramentas SeedServer, para agrupamento de sequências protéicas homólogas e U-MAGE, para propagação de ontologia funcional

Rafael Lucas Muniz Guedes

Belo Horizonte, Abril de 2013.
Instituto de Ciências Biológicas - UFMG
Bioinformática

Instituto de Ciências Biológicas

TESE DE DOUTORADO

Desenvolvimento das ferramentas SeedServer, para agrupamento de sequências protéicas homólogas e U-MAGE, para propagação de ontologia funcional.

Rafael Lucas Muniz Guedes

Tese de Doutorado apresentada ao Programa de Doutorado em Bioinformática, Departamento de Bioquímica e Imunologia da UFMG como requisito parcial para obtenção do título de Doutor em Bioinformática.

Orientador: Prof. Dr. José Miguel Ortega

Ata de Aprovação

Desenvolvimento das ferramentas SeedServer, para agrupamento de sequências protéicas homólogas e U-MAGE, para propagação de ontologia funcional.

Rafael Lucas Muniz Guedes

Tese defendida e aprovada pela banca examinadora constituída por:

Prof. Dr. José Miguel Ortega - Orientador
Universidade Federal de Minas Gerais

Profa. Dra. Daniella Castanheira Bartholomeu
Universidade Federal de Minas Gerais

Dr. Francisco Pereira Lobo
EMBRAPA

Prof. Dr. Guilherme Corrêa de Oliveira
Instituto René Rachou FIOCRUZ

Profa. Dra. Raquel Minardi
Universidade Federal de Minas Gerais

Agradecimentos

Agradeço a todos que de alguma forma contribuíram para o desenvolvimento dessa tese, em especial ao mentor Prof. Dr. J. Miguel Ortega com seus diversos e-mails noturnos, aos meus pais Antônio e Rosa Guedes, irmãos Paulo, Pedro e Thiago, aos atuais e antigos membros do Laboratório de Biodados e a atenção e carinho de Mariana Boroni.

RESUMO

Com o avanço das tecnologias de sequenciamento, a organização das bases de dados secundárias é uma contribuição importante, nas quais o conhecimento existente é organizado com base em informações biológicas. Agrupamento de genes homólogos e a atribuição de termos de ontologia funcional envolvendo genes já conhecidos são formas de acelerar a análise de dados de novos sequenciamentos. Neste trabalho relatamos o desenvolvimento do SeedServer, U-MAGE e aplicações. A integração de bancos de dados e programa agrupador de proteínas homólogas capazes de incluir sequências provenientes de genomas incompletos, conjuntamente com metodologias de validação e comparação de estruturas secundárias resultou no desenvolvimento da ferramenta SeedServer. Grupos de homólogos criados a partir de sequências de interesse do usuário são gerados com auxílio de uma interface *web* onde também é possível descarregar as sequências agrupadas, obter relatórios com taxonomia completa e estimar a origem do gene em questão através da determinação do ancestral comum mais recente. O programa SeedServer, após ser testado e avaliado foi então utilizado em um estudo da heterotrofia de aminoácidos através da formação de grupos de homólogos das enzimas presentes nas vias biossintéticas de aminoácidos essenciais, demonstrando um quadro denominado de Grande Deleção Genômica em diferentes grupos de eucariotos e procariotos. A esse evento pode se suceder a perda da capacidade de assimilação de nitrogênio, componente essencial na formação dos aminoácidos. Estudos filogenéticos mostraram uma maior taxa de mutação dentre as enzimas remanescentes de vias incompletas quando comparadas com outras de vias completas. Adicionalmente, para melhorar a qualidade da anotação funcional de sequências protéicas, foi criada a ferramenta denominada U-

MAGE (*UniRef50 Matrices for Annotation of Gene Ontology Entries*) que utiliza como base de propagação de termos de ontologia funcional, matrizes de cobertura entre sequências de mesmo UniRef50. O U-MAGE demonstrou uma melhora qualitativa significativa na anotação de ontologia funcional de diversos organismos. As duas ferramentas SeedServer e U-MAGE contribuem para a aceleração da propagação de informação de proteínas conhecidas, um desafio atual imposto à Bioinformática para fazer frente à intensa produção de novas sequências.

ABSTRACT

With advances in sequencing technologies, an important contribution is the organization of secondary databases, where existing knowledge is organized based on biological information. Grouping of homologous genes and assigning terms of functional ontology involving already known genes are ways to speed the analysis of data from new sequencing. We report the development of SeedServer, U-MAGE and applications. The integration of databases and a program capable of clustering homologous proteins including sequences derived from incomplete genomes, together with validation methods and comparison of secondary structures resulted in the development of SeedServer tool. Groups of homologous sequences chosen from user interest are generated with the aid of a web interface where you can also download the grouped sequences, get taxonomy reports and estimate the origin of the gene in question by determining the lowest common ancestor. The program SeedServer after being tested and evaluated was then used in a study of amino acid heterotrophy by forming groups of homologous enzymes present in the essential amino acids biosynthetic pathways, showing a scenario called the Great Genomic Deletion in different groups of eukaryotes and prokaryotes. Following that event may be the loss of assimilative capacity of nitrogen, an essential component in the formation of amino acids. Phylogenetic studies showed a higher rate of mutation among the enzymes remaining in incomplete pathways when compared with others from complete pathways. Additionally, to improve the quality of functional annotation of protein sequences, we created the tool called U-MAGE (UniRef50 Matrices for Annotation of Gene Ontology Entries) that uses as the basis of propagation of functional ontology terms the coverage between sequences within a

UniRef50 organized in matrices. The U-MAGE demonstrated a significant qualitative improvement in functional ontology annotation of various organisms. Both tools SeedServer and U-MAGE contribute to the acceleration of the information spread from known proteins, a challenge to the current Bioinformatics to face the intense production of new sequences.

I. Sumário

I. Sumário 1

II. Índice de figuras 4

III. Índice de tabelas 6

IV. Abreviações e siglas 7

1. Introdução 8

1.1. Homologia e conceitos 8

1.2. Métodos para agrupamentos de sequências homólogas 11

1.3. Bases de dados biológicos 13

1.3.1. *Kyoto Encyclopedia of Genes and Genomes* - KEGG 13

1.3.2. UniProt 14

1.3.3. *Protein ANalysis THrough Evolutionary Relationships* - PANTHER 15

1.4. Ontologia, *Gene Ontology* (GO) e *Gene Ontology Annotation* (GOA) 15

1.5. Aminoácidos essenciais e não-essenciais (estudo de caso) 18

2. Justificativa 20

3. Objetivos 20

4. Acessibilidade 21

5. Metodologia 22

5.1. Aplicadas ao desenvolvimento do SeedServer 22

5.1.1. Bancos de dados 22

5.1.2. Programas 26

5.1.3. Busca de homólogos 28

5.1.4. Validação e comparação da estrutura secundária 29

5.1.5. Relatório taxonômico final 32

5.1.6. *Web Services* 32

5.1.7. Página *web* SeedServer 33

5.2. Metodologia para estudos de aminoácidos essenciais 34

5.2.1. Bancos de Dados e programas..... 34

5.2.2. Procedimento de inspeção das vias 34

5.2.3. Análises de distâncias filogenéticas..... 35

5.3. Propagação dos termos de ontologia funcional protéica 35

5.3.1. Página *web* U-MAGE 38

6. Resultados 39

6.1. SeedServer 39

6.1.1. Aminoacil-tRNA sintetases..... 43

6.1.1.1. Agrupamentos..... 43

6.1.1.2. Casos de rejeição na validação PSI-BLAST 50

6.1.1.3. Correlação entre número EC e validação PSI-BLAST..... 51

6.1.1.4. Correlação entre família PANTHER, número EC e validação PSI-BLAST..... 56

6.1.1.5. Enriquecimento taxonômico..... 60

6.1.2. Metilaspartato mutase e subgrupos KO 64

6.1.3. Vitamina B7 e anotação de sequências metagenômicas 66

6.1.4. Via regulatória do desenvolvimento pré-embrionário e inferência de LCA..... 67

6.1.5. LCA e estrutura secundária 69

6.1.6. Contribuição de genomas não completos e componentes SeedServer 73

6.1.7. SeedServer em interface *web* 74

6.2. SeedServer e o estudo da extinção de vias metabólicas 81

6.2.1. Aminoácidos essenciais 81

6.2.2. Via biossintética de lisina 86

6.2.3. Auxotrofia para nitrogênio 89

6.2.4. Enzimas remanescentes nas vias de AEs 93

6.3. Propagação de termos de ontologia funcional baseada em matrizes UniRef50.... 98

6.3.1. U-MAGE em interface *web* 101

7. Discussão 106

8. Considerações finais 111

9. Referências 112

10. Produção científica durante o Doutorado 118

10.1. Artigos científicos publicados em revistas internacionais 118

10.2. Capítulos de livros 118

10.3. Trabalhos apresentados em congressos 118

11. Anexo I 122

12. Anexo II 136

II. Índice de figuras

Figura 1: Homólogos e suas subdivisões 10

Figura 2: Ontologia de processos biológicos aplicada ao estudo da função de genes envolvidos com câncer 17

Figura 3: Mapa físico do banco contendo as tabelas utilizadas pelo SeedServer. 25

Figura 4: Esquema ilustrando a utilização dos *Web Services* BOWS e LCAWS através do SeedServer. 33

Figura 5: Mapa físico do banco contendo as tabelas utilizadas pelo U-MAGE 38

Figura 6: Fluxograma com procedimento SeedServer completo.. 43

Figura 7: Árvores filogenéticas contendo *Seeds* de archaeas para sub-grupos de aminoacil-tRNA sintetases divididos pelo Seed Linkage. 45

Figura 8: Diagrama de Venn representando as categorias existentes ao final do agrupamento SeedServer. 46

Figura 9: Árvore filogenética das 51 metilaspartato mutases presentes no K01846..... 65

Figura 10: Reposicionamento do ancestral comum mais recente (LCA) com uso do SeedServer. 68

Figura 11: Correlação da sobreposição de estruturas secundárias dada pelo valor SOV com distância taxonômica. 70

Figura 12: Distribuição em classes distintas de valores SOV obtidos em três níveis taxonômicos em razão do total de dados 72

Figura 13: Participação de proteínas provenientes de genomas com diferentes níveis de sequenciamento 73

Figura 14: Participação de proteínas provenientes das diferentes porções do recrutamento SeedServer 74

Figura 15: Página *web* principal desenvolvida para disponibilização do SeedServer... 75

Figura 16: Página *web* para visualização dos resultados SeedServer..... 77

Figura 17: Página *web* contendo relatório da presença/ausência das proteínas em grupos taxonômicos no mesmo nível e um nível inferior ao LCA..... 79

Figura 18: Ausência de proteínas envolvidas na resistência a seca em plantas em diversos grupos taxonômicos. 80

Figura 19: Representação esquemática para presença/ausência das enzimas biossintéticas dos nove aminoácidos essenciais e dos não essenciais serina e glicina. .. 86

Figura 20: Representação esquemática para presença/ausência das enzimas de biossíntese de lisina..... 89

Figura 21: Representação esquemática para presença/ausência de glutamato desidrogenases.. 92

Figura 22: Árvores filogenéticas de A: acetolactato sintase (VIL1) e B: um grupo de alanina-glioxilato, serina-glioxilato e serina-piruvato transaminases (G1) 96

Figura 23: Distâncias relativas de enzimas das vias de aminoácidos essenciais e não essenciais de metazoários para homólogos presentes em fungos e plantas. 97

Figura 24: Enriquecimento quantitativo e qualitativo da propagação U-MAGE para 82 organismos da base de dados PANTHER. 100

Figura 25: Página *web* principal desenvolvida para disponibilização do U-MAGE.. 102

Figura 26: Propagação U-MAGE para a proteína humana O60674..... 105

III. Índice de tabelas

Tabela 1: Exemplo de tabela final com resultados SeedServer..... 31

Tabela 2: Matriz hipotética contendo dados de cobertura de membros UniProtKB de mesmo UniRef50..... 37

Tabela 3: Resultados do agrupamento SeedServer para as aminoacil-tRNA sintetases.....49

Tabela 4: Correlação entre número EC e validação PSI-BLAST..... 55

Tabela 5: Correlação entre família PANTHER, número EC e validação PSI-BLAST.....58

Tabela 6: Correlação entre famílias e subfamílias PANTHER e validação PSI-BLAST.....61

Tabela 7: Enriquecimento taxonômico como o recrutamento SeedServer..... 63

IV. Abreviações e siglas

aaRS - aminoacil-tRNA sintetase

AE - Aminoácidos Essenciais

BBH - *Bidirectional Best Hit* (melhor alinhamento recíproco)

BLAST - *Basic Local Alignment Search Tool*

BOWS – *Bioinformatics Open Web Services*

COG - *Clusters of Orthologous Groups*

EC - *Enzyme Comission*

GDG - Grande Deleção Genômica

GO - *Gene Ontology* (Ontologia Gênica)

KEGG - *Kyoto Encyclopedia of Genes and Genomes*

KO - KEGG *Orthology* (KEGG Ortologia)

LCA - *Lowest Common Ancestor* (ancestral comum mais recente)

LCAWS - *Lowest Common Ancestor Web Service*

Mut - Metilaspartato mutase

PANTHER - *Protein ANalysis THrough Evolutionary Relationships*

SOV - *Secondary Structure OVerlap* (sobreposição de estrutura secundária)

UEKO - *UniRef50 Enriched* KEGG *Orthology*

U-MAGE – *UniRef90 Matrices for Annotation of Gene Ontology Entries*

WS - *Web Service*

WSDL - *Web Services Description Language*

1. Introdução

Uma importante contribuição da Bioinformática para a análise rápida de novos sequenciamentos é a organização da informação já conhecida em bases de dados sendo um importante recurso apresentado pelas bases de dados de genes homólogos, nas quais os mesmos apresentam-se organizados por categorias funcionais. Algumas vezes, genes de interesse de um investigador não são encontrados nessas bases de dados ou a informação esta incompleta. O trabalho desenvolvido por nosso grupo atua nessa lacuna. Para uma melhor compreensão do tema, revisamos inicialmente o conceito de homologia, os métodos atuais para criação de grupos de homólogos, as bases de dados de homólogos disponíveis – ou que se referem a elas – e os agrupamentos por ontologia gênica que consistem em um recurso análogo. Introduzimos também nessa seção o tema "aminoácidos essenciais", pois aplicamos as ferramentas desenvolvidas em um estudo de caso focado na origem da heterotrofia de aminoácidos.

1.1. Homologia e conceitos

Os genomas dos organismos estudados hoje são como retratos parciais, agrupados e comparados na tentativa de elucidar a cadeia de eventos contínuos ocorridos durante o processo evolutivo, tais como mutações, rearranjos, duplicações, inserções e deleções. Tais variações genéticas, com susceptibilidade de ocorrência em regiões não codificantes como codificantes, são uma das diretrizes que moldam o ganho ou perda de funções metabólicas. Dois ou mais genes separados ao longo desse processo de evolução, descendentes de um ancestral comum, são intitulados genes homólogos e podem ainda desempenhar o mesmo papel biológico. Esses genes encontram-se sob pressões seletivas e embora acumulem mutações ao longo

de suas existências, guardam ainda similaridades. A similaridade da cadeia de aminoácidos e/ou estrutural é tida como forte indicativo da manutenção da função. Por esse motivo, técnicas de agrupamento que correlacionam proteínas de função desconhecida a outras já descritas na literatura são valiosas e muito utilizadas para inferência e propagação da informação biológica, bem como para estudos de evolução gênica, de genes taxonomicamente exclusivos e da genômica comparativa.

Como proposto por Sonnhammer e Koonim [1] o conceito de homologia mais difundido e aceito pode ser subdividido nas seguintes categorias principais: (i) Ortólogo: genes homólogos que sofreram um evento de especiação; (ii) Parálogos: genes homólogos resultantes de um evento de duplicação na mesma espécie; (iii) In-Parálogos: Parálogos resultantes de um evento de duplicação ocorrido após um evento de especiação e por fim (iv) Out-Parálogos: Parálogos resultantes de um evento de duplicação ocorrido anteriormente a um evento de especiação.

A Figura 1 organiza espacialmente e temporalmente as divisões citadas acima. Cada espécie é representada por um formato distinto, círculo, cruz, quadrado e losango enquanto os genes por letras e números. Uma espécie ancestral, representada por um retângulo, passa inicialmente por um processo de especiação (Especiação 1) seguido de um evento de duplicação (Duplicação 1) do gene A em questão. Um segundo evento de especiação se segue (Especiação 2) e uma das cópias de A passa então por outro evento de duplicação (Duplicação 2). Consideremos que B1 e C1 mantêm a mesma sintenia da cópia A, então A, B1 e C1 são exemplos de genes ortólogos. Na espécie representada pelo quadrado, B1 e B2 são parálogos, assim como C2a e C2b são parálogos na espécie representada pelo losango.

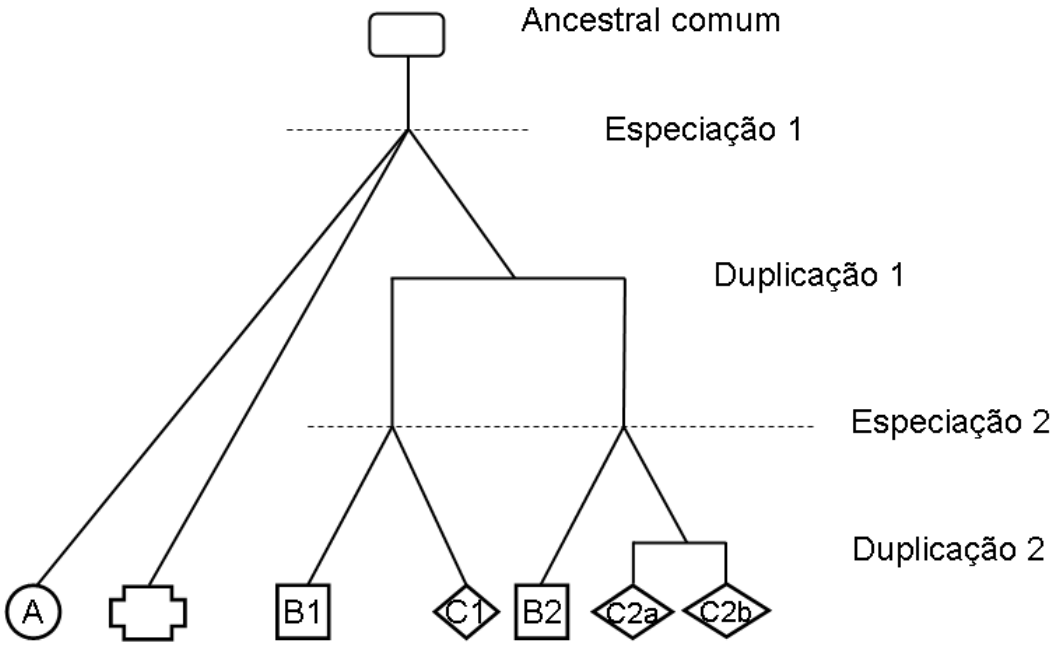


Figura 1: Homólogos e suas subdivisões. Espécies distintas estão representadas por círculo, cruz, quadrado e losango; o ancestral está representado pelo retângulo; os genes, representados por letras e números. Cruz vazia representa deleção do gene. Linha tracejada: evento de especiação; forquilha reta: evento de duplicação gênica. Ver o texto para explicação.

Note que a Especiação 2 é subseqüente à Duplicação 1. Eventos subseqüentes de especiação tendem a favorecer neo-funcionalização (surgimento de função totalmente nova) de genes duplicados, os quais acabam por formar novas famílias gênicas. É possível que os genes B1 e B2 constituam já famílias gênicas diferentes, e a designação mais específica proposta nesse caso é out-parálogos. Em comparação, os genes C2a e C2b podem ser designados in-parálogos, os parálogos propriamente ditos, pois se trata de duplicação na mesma linhagem. Na formação de grupos de homólogos é crítico inserir out-parálogos em grupos distintos. Uma perda ou deleção gênica está representada por uma cruz vazia ocorrida junto ao primeiro evento de especiação. O estudo de perdas gênicas fica mais possível de ser estudado à medida que mais genomas completos ficam disponíveis.

A criação de parálogos é tida como uma fonte na geração de novos genes, uma vez que enquanto uma cópia desempenha seu papel no metabolismo a outra está livre de pressões seletivas e, portanto susceptível a alterações. Dessa forma os parálogos podem sofrer sub-funcionalização (dividir funções pré-existentes) ou mesmo neo-funcionalização, como mencionado anteriormente.

Diferentes parâmetros relativos à similaridade e filogenia das sequências são então empregados para agrupar genes dessas categorias, mas de maneira geral, em um grupo de homólogos objetiva-se a inclusão de ortólogos e in-parálogos, uma vez que out-parálogos são mais propensos a terem sofrido alteração de função biológica e comumente tidos como não ortólogos e sim membros de famílias gênicas distintas como, por exemplo, é caso dos diferentes tipos de histonas e globinas.

1.2. Métodos para agrupamentos de sequências homólogas

Devido a importantes utilidades em anotações funcionais e estudos evolutivos, diferentes metodologias foram propostas para geração e disponibilização de bancos de dados de grupos de homólogos. Geralmente, tais métodos exigem a inclusão de apenas genomas completos para análises. Comparar diversos proteomas, par a par, extraindo o melhor alinhamento recíproco ("*Bidirectional Best Hit*" – BBH) em uma comparação de todas as proteínas contra todas utilizando o algoritmo BLAST [2] é uma forma simples e eficiente de inferir homologia, quando um número relativamente pequeno de genomas são considerados, como ocorre no algoritmo InParanoid [3]. Apesar de se limitar a pares de organismos é um algoritmo capaz de distinguir in-parálogos de out-parálogos, dadas algumas regras internas como não agrupar genes de um mesmo organismo se o escore for inferior ao do alinhamento com um gene do outro organismo, além de não necessitar de árvores filogenéticas.

A comparação par a par também pode ser estendida de forma a envolver mais de dois organismos, através da formação de melhores cruzamentos triangulares [4–6], porém esses métodos são mais sensíveis a erro em casos de deleções gênicas - uma vez que a relação entre homólogos está perdida - e a proteínas com múltiplos domínios. Para minimizar este problema foi desenvolvido um aprimoramento do método BBH que leva em consideração as deleções gênicas e distâncias evolucionárias [7]. Há também técnicas como as aplicadas na base de dados do OrthoMCL [8] que fazem uso de modelos de Markov para refinar o agrupamento final.

Partindo-se de grupos de homólogos pré-computados, alguns métodos constroem árvores filogenéticas para distinção entre ortólogos e parálogos [9], porém para melhor detectar eventos de duplicação e especiações gênicas, métodos híbridos (similaridade de sequência mais árvores filogenéticas) foram desenvolvidos. Mesmo contendo poucas dezenas de genomas eucarióticos, Ensembl Compara [10] é uma fonte atualizada de grupos de homólogos com a vantagem de serem menos influenciáveis por deleções gênicas enquanto os grupos PHOGs [11] lidam bem com proteínas multi-domínios. Já o TreeFam é uma base manualmente curada [12]. No entanto, quando comparados entre si, métodos baseados em filogenia mostram-se menos eficientes do que os algoritmos de similaridade de sequência [13].

Com a crescente facilidade em sequenciamento de genomas, manter a atualização, com buscas intensivas de BLAST, dos grandes bancos de dados de grupos de homólogos [14] deve se tornar computacionalmente proibitivo. Nesse sentido, nosso grupo recentemente desenvolveu uma ferramenta capaz de criar grupos de homólogos sob demanda a partir de sequências protéicas de interesse, com a grande vantagem de incluir informações provenientes de genomas

incompletos [15]. Desenvolvemos também uma metodologia para enriquecer bases de dados já existentes como o COG - *Clusters of Orthologous Groups* - [16] a aplicamos ao enriquecimento de grupos KEGG Orthology (KO) [17], detalhado posteriormente (dados não publicados desenvolvidos durante o doutoramento em bioinformática do Dr. Gabriel da Rocha Fernandes pela Universidade Federal de Minas Gerais em 2011) [18].

1.3. Bases de dados biológicos

É esperado que bancos de dados biológicos que se propõem a viabilizar dados de um determinado campo de estudo à comunidade científica, como por exemplo, grupos de sequências homólogas, independentemente do algoritmo utilizado para defini-los, respeitem algumas características fundamentais como: confiabilidade, atualizações periódicas, permitir acesso aos dados pelos usuários, disponibilidade pela internet e não somente via instalação local, além de interação com outras bases de dados para enriquecer a qualidade da informação.

Nos tópicos dessa seção serão abordados os bancos de dados utilizados como referência no presente trabalho que obedecem a essas regras.

1.3.1. *Kyoto Encyclopedia of Genes and Genomes* - KEGG

A base de dados KEGG é uma fonte de informação consistente, organizado em diferentes módulos como KEGG *Genome* (contém informações de todos os genomas completos disponíveis), KEGG *Medicus* (correlaciona informações entre drogas e doenças humanas), mas iremos ressaltar aqui os módulos KEGG *Orthology* (KO) e KEGG *Pathway*.

O KO apresenta uma relação de mais de 14 mil grupos de homólogos de sequências provenientes de genomas completos dos três grandes domínios de seres vivos, eucariotos, procariotos e archaeas. Esses grupos são formados de forma automática e manual, ambas através da análise de alinhamentos BBH que são processados e armazenados para todos os genes. A cada KO é designado um número K (exemplo: K00001 - álcool desidrogenase) e sempre que a função enzimática é conhecida, associa-se também o código enzimático padronizado e hierárquico estabelecido pela IUBMB (*International Union of Biochemistry and Molecular Biology*) conhecido como número EC – *Enzyme Commission*. No exemplo anterior, álcool desidrogenase possui o EC: 1.1.1.1.

Para organizar os grupos KO criou-se o KEGG *Pathway*, onde é possível inspecionar diversas vias e reações metabólicas, evidenciando-se pontualmente a presença ou ausência de cada função enzimática em todos os organismos representados.

1.3.2. UniProt

A base de dados UniProt [19] é uma referência para sequências protéicas, possui informações funcionais e estruturais de alta qualidade, livre acesso e fácil navegação.

No UniProt encontramos dois módulos principais, o UniProtKB (*Protein Knowledgebase*) onde de fato se encontra toda a anotação das sequências protéicas e o UniRef (*UniProt Reference Clusters*) [20] que agrupa as sequências do UniProt em três níveis de identidade, 50%, 90% e 100% onde cada membro de um grupo deve ter, além da respectiva identidade, 80% de cobertura em relação a sequência

escolhida como representante de cada grupo, que geralmente é a que possui maior número de resíduos de aminoácidos.

Por sua vez, o UniProtKB também pode ser divido em dois módulos, sequências que passaram por curadoria manual (Swiss-Prot) e aquelas anotadas por métodos automáticos (TrEMBL).

1.3.3. *Protein ANalysis THrough Evolutionary Relationships* - PANTHER

O PANTHER [21] é uma base de dados manualmente curada por especialistas que organiza diversas famílias protéicas e suas respectivas subfamílias, cada qual com anotação funcional e árvores filogenéticas contendo o recentemente atualizado número de 82 organismos, incluindo também eucariotos, procariotos e archaeas. Esses dados são úteis para estudos de eventos evolutivos como duplicações gênicas e especiações além de possibilitar a classificação de sequências em famílias protéicas para análises de larga escala.

1.4. Ontologia, *Gene Ontology* (GO) e *Gene Ontology Annotation* (GOA)

O termo 'ontologia' em uma definição mais abrangente significa uma forma de organizar, definindo classes e hierarquias aos aspectos fundamentais de uma determinada entidade em estudo.

Como uma tentativa de padronizar dados de anotação gênica surgiu o *Gene Ontology* (GO), que estrutura a informação utilizando vocabulários controlados em hierarquias contendo desde termos mais abrangentes como 'atividade catalítica' até

termos mais específicos como ‘receptor transmembrana com atividade tirosina quinase’.

Três grandes hierarquias foram criadas de forma a abranger todas as características de uma anotação, são elas: (i) - Função Molecular, (ii) - Processo Biológico e (iii) - Componente Celular.

A partir de um conjunto de dados, por exemplo, transcriptômicos ou proteômicos, é possível caracterizar por comparação de similaridade com um banco de dados as funções metabólicas presentes. Para isso, existem ferramentas [22–24] capazes de estabelecer essa correlação nos diversos níveis da hierarquia, com suporte estatístico, e gerar figuras bastante informativas como a Figura 2, que auxilia o pesquisador na interpretação biológica da função de genes envolvidos com câncer.

É importante ressaltar que todas as bases de dados citadas nessa e na seção anterior são facilmente integradas. Identificadores protéicos UniProtKB são associados às sequências presentes em cada grupo KO, bem como a famílias protéicas PANTHER. Por sua vez, identificadores de ontologia são designados por especialistas, tanto aos identificadores das famílias PANTHER, quanto aos UniProtKB, através de um consórcio chamado *Gene Ontology Annotation* (GOA) que gera anotações de alta qualidade.

No entanto, esse trabalho manual de atribuição de termos de ontologia acaba criando um viés na amostragem desses bancos de dados uma vez que possuem principalmente organismos modelo como alvo, e. g. *Homo sapiens* e *Mus musculus*. Com isso, organismos menos estudados acabam permanecendo sub-representados mesmo com alguns esforços existentes de anotação automática.

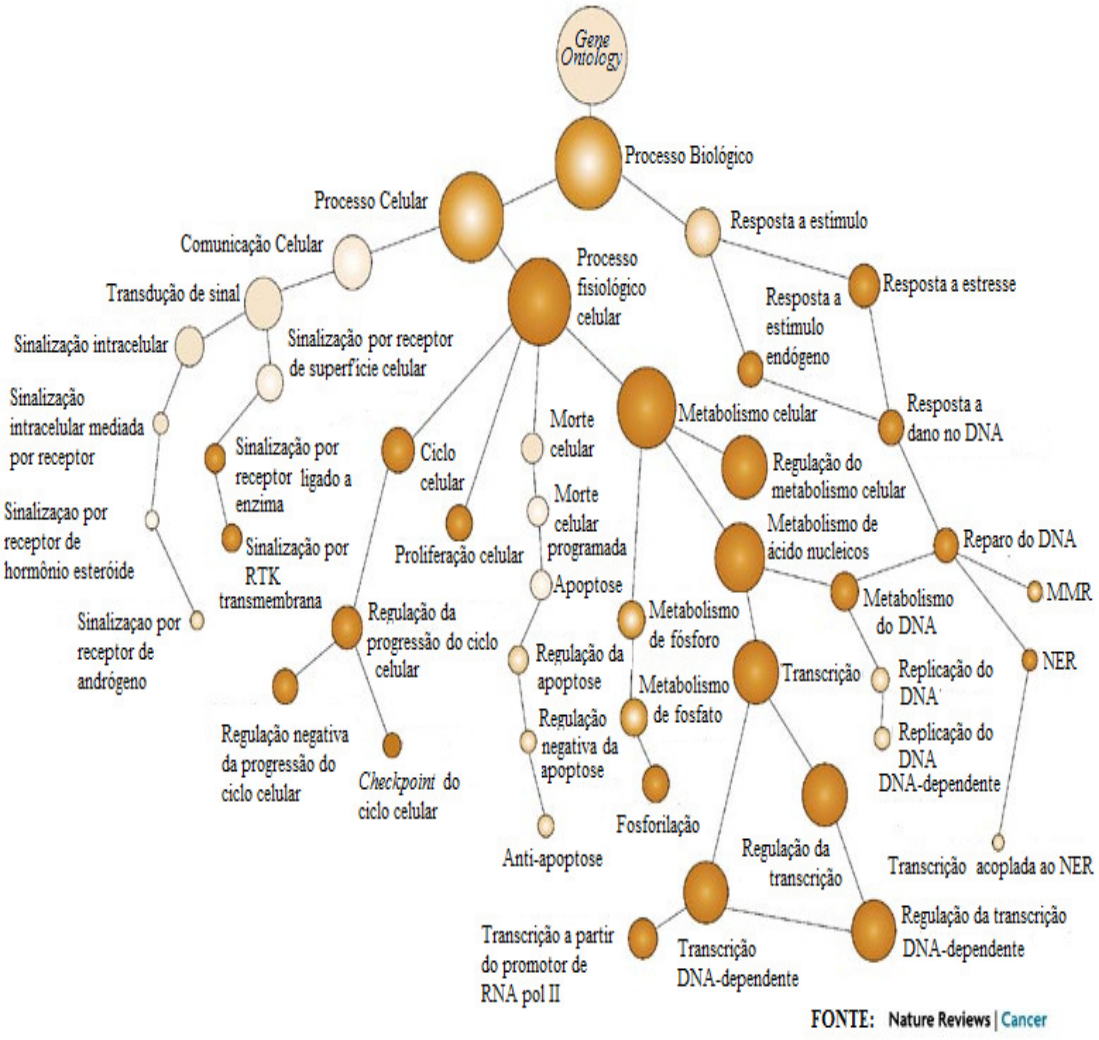


Figura 2: Ontologia de processos biológicos aplicada ao estudo da função de genes envolvidos com câncer. O tamanho dos círculos representa a quantidade de genes associados e as cores mais escuras um maior suporte estatístico (*p-value*) para cada categoria. Traduzido de: Hu *et al*. Nature Reviews Câncer **7**, 23-24 (Jan – 2007).

Por exemplo, a proteína humana P33897 (membro 1 da subfamília D de proteínas cassete com domínio de ligação a ATP) foi uma das escolhidas em um trabalho recente focado no estudo das proteínas presentes no peroxissomo [25]. Já a proteína G3QS26 de outro primata *Gorilla gorilla*, com 99% de identidade BLAST e mesmo

tamanho, continua sem anotação de termos envolvidos no trabalho citado, como por exemplo o GO:0005325 (atividade transportadora de acil-CoA em peroxissomos).

1.5. Aminoácidos essenciais e não-essenciais (estudo de caso)

Além das aplicações mais comuns já citadas como inferência de função gênica, os grupos de homólogos são também fontes importantes para estudos de deleções gênicas e perda de função de vias metabólicas.

Nesse sentido, foram escolhidas como objeto de estudo de parte desse trabalho as vias biossintéticas de aminoácidos essenciais (AEs).

O anabolismo de aminoácidos é responsável por cerca de 20% da energia gasta pelas células durante a síntese protéica [26, 27]. A necessidade de ingestão dos aminoácidos dos quais somos incapazes de sintetizá-los *de novo*, portanto denominados de aminoácidos essenciais, é de extrema importância e a composição corporal humana de AEs foi estimada em 22 mg/kg e, de nitrogênio, em 3 mg/kg [28, 29]. Estudos mais recentes revelam que essa dependência pode ser até cinco vezes maior do que previamente estimado além de quantificar a necessidade individual aos nove AEs humanos (histidina, fenilalanina, triptofano, valina, isoleucina, leucina, lisina, metionina e treonina) [30]. Importante salientar que todos os aminoácidos também nos provêem de uma fonte de nitrogênio orgânico.

É de conhecimento geral que plantas e fungos possuem todas as enzimas necessárias para a síntese *de novo* dos 20 aminoácidos e que deleções de alguns desses genes ocorreram em nossos ancestrais ao logo da evolução. Os motivos exatos de como tais organismos tornaram-se aptos a sobreviver e competir com outras formas de vida, sem essas habilidades essenciais, mantém-se como uma pergunta em aberto.

O padrão de perda e retenção das enzimas de biossíntese de alguns aminoácidos foi analisado para alguns protistas e metazoários por Payne e Loomis [31]. Eles verificaram um conjunto de AEs comuns entre animais e protistas, indicativo de que uma Grande Deleção Genômica (GDG) ocorreu nos ancestrais desses organismos. Curiosamente, a maioria das enzimas de biossíntese dos aminoácidos essenciais mantidas é intermediária de outras vias metabólicas, como biossíntese do anel de purinas e metabolismo do nitrogênio [31]. Algumas são mantidas pelo papel na degradação específica de alguns aminoácidos, o que é coerente com a necessidade de aproveitamento de nitrogênio orgânico.

A utilização de bancos de dados de homólogos atualizados, com proteínas de organismos de taxonomias diversas, poderia ajudar a construir um cenário mais completo a respeito desses eventos de deleção e do surgimento da auxotrofia de aminoácidos.

2. Justificativa

Grupos de proteínas homólogas são importantes fontes para propagação da informação biológica, bem como inferência de deleções gênicas em genomas. No entanto é cada vez mais difícil manter a atualização com o crescente número de novas sequências depositadas diariamente nos bancos de dados, fruto do desenvolvimento das novas tecnologias de sequenciamento a baixos custos. Dessa forma, faz-se necessário o desenvolvimento de novas ferramentas capazes de (i) Lidar com esse volume de informação, agrupando com eficiência qualquer grupo de sequências homólogas, incluindo se possível processamento sob demanda e (ii) Ajudar na propagação da anotação manual de algumas poucas sequências para outras estritamente relacionadas de forma segura e automática para enriquecer a qualidade da informação associada às sequências e seus homólogos.

3. Objetivos

- Integrar programas e bancos de dados para estudos de proteínas homólogas em uma ferramenta denominada SeedServer.
- Validar a ferramenta SeedServer.
- Utilizar o SeedServer para estudo da deleção de genes e vias biossintéticas.
- Desenvolver um servidor *web* de fácil utilização onde os usuários possam utilizar o SeedServer sem necessidade de instalação local.
- Desenvolver uma ferramenta capaz de propagar anotação funcional de proteínas de forma segura, denominada U-MAGE.
- Validar a ferramenta U-MAGE.

- Desenvolver um servidor *web* de fácil utilização onde os usuários possam utilizar o U-MAGE sem necessidade de instalação local.

4. Acessibilidade

Os artigos publicados durante o desenvolvimento desse trabalho podem ser obtidos em:

- **Amino acids biosynthesis and nitrogen assimilation pathways: a great genomic deletion during eukaryotes evolution.** **Guedes RL**, Prosdocimi F, Fernandes GR, Moura LK, Ribeiro HA, Ortega JM. BMC Genomics. 2011 Dec 22;12 Suppl 4:S2. PMID: 22369087.

  - http://www.biomedcentral.com/1471-2164/12/S4/S2 e Anexo 1.

  Para árvores filogenéticas adicionais desse estudo:

  - http://biodados.icb.ufmg.br/AEs/

- **Preimplantation development regulatory pathway construction through a text-mining approach**: Donnard E, Barbosa-Silva A, **Guedes RL**, Fernandes GR, Velloso H, Kohn MJ, Andrade-Navarro MA, Ortega JM. BMC Genomics. 2011 Dec 22;12 Suppl 4:S3. PMID: 22369103.

  - http://www.biomedcentral.com/1471-2164/12/S4/S3 e Anexo 2.

SeedServer e U-MAGE estão respectivamente disponíveis em:

- http://pinguim.fmrp.usp.br/cenabid/form_SS.html
- http://pinguim.fmrp.usp.br/u-mage/form_u-mage.html

Os WSDLs BOWS e LCAWS, descritos na seção 5.1.6 e um guia para utiliza-los estão respectivamente disponíveis em:

- http://pinguim.fmrp.usp.br:8080/BOWS/services/BOWS?wsdl
- http://merengue.icb.ufmg.br:8080/BioToolsService/services/lca?wsdl
- http://biodados.icb.ufmg.br/services/

5. Metodologia

5.1. Aplicadas ao desenvolvimento do SeedServer

Para a criação da ferramenta SeedServer, diferentes programas e bases de dados foram integrados sob o comando de um programa principal desenvolvido na linguagem de programação PERL [www.perl.org].

5.1.1. Bancos de dados

- **Sequências Protéicas:** O banco de dados protéico disponível para consulta de todas as partes constituintes do procedimento SeedServer é formado por sequências provenientes da base UniProtKB no formato FASTA, com remoção das identificadas como fragmentos restando portanto somente aquelas designadas como *Complete*. Para essa exclusão um programa na linguagem PERL foi criado e utilizado para encontrar quando presente o termo "*fragment*" adicionado na descrição da sequência FASTA. Os arquivos de sequências uniprot_sprot.fasta.gz e uniprot_trembl.fasta.gz podem ser obtidos em: [ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete].
- **Informações secundárias:** A partir dos dados da fonte mencionada acima foi possível extrair informações relevantes atribuídas às proteínas como descrição, presença de curadoria manual SwissProt, número EC, número de resíduos de aminoácidos (através de um programa PERL criado para contagem do número de letras de aminoácidos por sequência nos arquivos FASTA) e então armazenados em tabela (Figura 3), em

banco local MySQL [http://www.mysql.com]. A versão UniProtKB utilizada nesse trabalho para o avaliação do SeedServer foi UniProtKB_2012_02 [ftp://ftp.uniprot.org/pub/databases/uniprot/previous_releases/release-2012_02/]. Quando do estudo de deleções em vias biossintéticas de aminoácidos essenciais, foi utilizada a versão UniProtKB_2010_10 [ftp://ftp.uniprot.org/pub/databases/uniprot/previous_releases/release-2010_10/].

- **Identificação taxonômica:** Outra tabela local MySQL foi criada contendo dados taxonômicos provenientes do serviço *Taxonomy* do *National Center for biotechnological Information* (NCBI) [ftp://ftp.ncbi.nih.gov/pub/taxonomy/]. Essa tabela contém identificadores taxonômicos numéricos para os diversos níveis hierárquicos como reino, filo e gênero para todos os organismos (Figura 3). Organismo é utilizado como a denominação técnica do alvo do sequenciamento, seja uma espécie, subespécie, linhagem, etc. Como cada sequência FASTA citada acima possui um identificador taxonômico específico para o organismo de origem é possível associar a taxonomia completa para cada proteína. A associação entre proteína e identificador taxonômico do organismo pode ser obtida em: [ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping]. Os níveis taxonômicos acima do nível de organismo foram extraídos da tabela taxsimple, produzida em outro projeto do laboratório e disponibilizada em [http://biodados.icb.ufmg.br/services/]. O nível de sequenciamento de cada genoma foi obtido

[ftp://ftp.ncbi.nih.gov/genomes/genomeprj_archive/] e armazenado na mesma tabela, com as distinções: (i) Genoma Completo, (ii) Genoma em Montagem, (iii) Genoma Incompleto e (iv) Genoma sem projeto de sequenciamento.

- **UniRef50:** *UniProt Reference Clusters 50* pode ser obtido em: [ftp://ftp.uniprot.org/pub/databases/uniprot/uniref/uniref50].

  Cada UniRef50 agrupa sequências com 50% de identidade e a partir do ano de 2013 um filtro extra de 80% de cobertura com a sequência representativa do grupo, que é geralmente a maior.

- **UniRef50 *Enriched* KEGG *Orthology* (UEKO):** UEKO é também uma tabela local MySQL (Figura 3) contendo os dados do enriquecimento de cada entidade KEGG *Orthology* (KO) disponível em sua última versão gratuita de 2012. Grupos UniRef50 ao qual cada proteína em um determinado KO pertence são consultados e o enriquecimento realizado seguindo filtros de qualidade descritos por Fernandes *et al.* [18], uma metodologia inicialmente utilizada para enriquecimento de agrupamentos protéicos do *Clusters of Orthologous Groups* (COG) [16]. Sucintamente, são realizados alinhamentos BLAST entre sequências originais do KO e as sequências presentes no UniRef90 de cada uma das originais, seguido de uma verificação de recobrimento (tamanho do alinhamento dividido pelo tamanho da sequência original) sendo exigido um mínimo de 50%. Essa metodologia é responsável por agregar sequências provenientes de genomas incompletos aos grupos KO e eleva o número de entradas protéicas em aproximadamente 300%, de 2.362.267 para 7.147.233.

- **Sequências metagenômicas:** Sequências no formato FASTA de diversos projetos públicos metagenômicos foram obtida em [ftp://ftp.ncbi.nih.gov/genbank/wgs/].

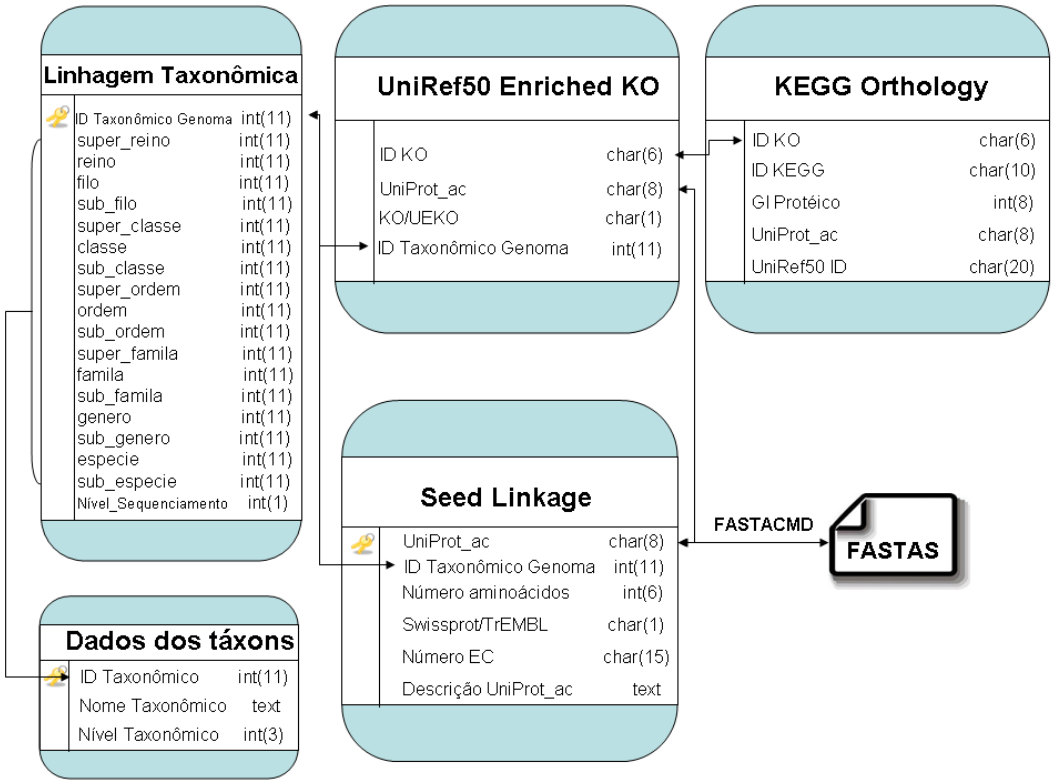


Figura 3: Mapa físico do banco contendo as tabelas utilizadas pelo SeedServer. As setas pretas indicam os identificadores que correlacionam as tabelas e arquivos. Todos os níveis de identificadores taxonômicos da tabela 'Linhagem Taxonômica' possuem nomes na tabela 'Dados dos táxons'. As chaves indicam quais são os identificadores únicos. O programa FASTACMD do pacote BLASTALL é utilizado para obter as sequências FASTA a partir de identificadores UniProtKB. ID: identificador. *Int*: capo reservado para números inteiros; *Char*: campo reservado para caracteres; *Text*: campo reservado para texto livre.

5.1.2. Programas

- Seed Linkage: É um programa de livre acesso desenvolvido em nosso grupo por Barbosa-Silva e colaboradores, capaz de criar grupos de homólogos a partir de uma única proteína, através de buscas exaustivas de similaridade de sequência com uso de BBH e é também capaz de usar genomas incompletos. O algoritmo começa alinhando uma sequência de um organismo de interesse com outras sequências de diversos genomas no banco de dados. As relações de BBH estabelecidas com outros genomas servem de parâmetro para as respectivas buscas de in-parálogos enquanto a melhor relação BBH obtida é utilizada para busca de in-parálogos no organismo de interesse. Maiores explicações de uso e instalação podem ser encontradas em [15]. Todos os experimentos foram realizados com parâmetros *default* e *Expected Value* ou *E-value* 10^{-10}.
- SeedServer: Também é um programa de livre acesso desenvolvido na linguagem PERL, v5.8.8. Um guia de instalação e uso local está disponível em: [http://biodados.icb.ufmg.br/SeedServer/]. O programa acopla e coordena várias rotinas como Seed Linkage e consultas a UEKO (acima).
- PSI-BLAST: Constituinte do pacote BLASTALL, o PSI-BLAST pode ser obtido em [ftp://ftp.ncbi.nih.gov/blast/executables/]. Os parâmetros utilizados foram: *E-value* 1e-05, *H-value* 1e-10 e 50 iterações. O módulo que executa esta aplicação foi desenvolvido na linguagem JAVA por Henrique A. L. Ribeiro (não publicado) e interrompe as iterações quando o auto-escore da sequência *query* fica abaixo de uma razão estabelecida (default 0,7).

- *Secondary Structure OVerlap* (SOV): A predição da estrutura secundária foi feita com PREDATOR [32] pela combinação de eficácia e tempo de processamento. O alinhamento das estruturas secundárias par a par (sequência de interesse contra recrutadas) foi feito com CLUSTALW [33] e finalmente a porcentagem de sobreposição calculada como descrito por [34] tomando o valor médio entre as comparações bi-direcionais. O módulo para este cálculo foi desenvolvido por Oto Coelho-Jr (não publicado).
- Análises filogenéticas: Alinhamentos múltiplos de sequências foram feitos com os parâmetros *default* do PRANKSTER [35], através de dois ciclos de alinhamento por assumirmos filogenia correta desconhecida. Este programa foi escolhido por ter demonstrado melhor desempenho que seus concorrentes na presença de eventos de deleções e inserções. MEGA5 [36] foi então utilizado para construir árvores filogenéticas com o algoritmo *neighbor-joining* [37] e 500 replicas de *bootstrap*. Todas as análises filogenéticas realizadas nesse estudo seguiram essa metodologia.
- *PANTHER HMM Scoring tool* - versão 1.03: Essa ferramenta [38] foi utilizada com parâmetros *default* para atribuir famílias protéicas PANTHER a sequências UniProtKB [38] sendo escolhido o melhor resultado.
- Seleção de melhores alinhamentos BLASTx com proteínas diferentes: Para casos de múltiplos alinhamentos BLASTx mapeados em diferentes regiões de uma mesma sequência *query*, um programa escrito em JAVA desenvolvido no Laboratório de Biodados por Lucas Santana dos Santos

(não publicado) foi utilizado para seleção dos melhores alinhamentos sem sobreposição.

5.1.3. Busca de homólogos

A implementação do processo de busca de homólogos com SeedServer começa com o carregamento de um ou mais identificadores de sequências protéicas completas provenientes da base de dados UniProtKB, daqui pra frente referidas como sequências *Seeds*. Carrega-se em seguida a taxonomia em nível de reino ou filo onde a busca irá ocorrer bem como parâmetros Seed Linkage e de validação PSI-BLAST (detalhada no próximo tópico). O programa Seed Linkage é executado e inicia a busca por homólogos. Em seguida o banco de dados UEKO é consultado e todas as sequências presentes no mesmo grupo UEKO da(s) *Seed*(s) são recrutadas e as informações devidamente sinalizadas e armazenadas em uma tabela MySQL. A coluna referente ao Seed Linkage mostra valor 2 para a *Seed*, 1 para recrutadas pelo Seed Linkage e 0 para não recrutadas por ele. A coluna referente ao UEKO registra valor 1 para sequências presentes no KO, 2 para presentes na porção enriquecida do UEKO, 0 para ausentes. Sempre que uma *Seed* não pertencer a nenhum grupo KO/UEKO é feito o procedimento denominado *UniRef50 Enriched-Seed* (UE-*Seed*) onde o grupo UniRef50 da *Seed* é consultado e é exigido que o tamanho do alinhamento BLAST seja superior a 50% to tamanho da *Seed*. A coluna referente ao UEKO registra o valor 3 para recrutadas pelo procedimento UE-*Seed*. São então agregadas ao grupo final informações como número de resíduos de aminoácidos, identificador taxonômico do organismo, se há curadoria manual SwissProt, o filo e a descrição protéica. O passo seguinte consiste na validação e comparação de estruturas secundárias, discutidos no próximo tópico.

5.1.4. Validação e comparação da estrutura secundária

O processo de validação de todas as proteínas recrutadas seja pelo Seed Linkage ou UEKO é baseado em matrizes PSI-BLAST. Em um primeiro passo é determinado um valor denominado auto-escore, que representa o valor dado em um alinhamento PSI-BLAST da *Seed* contra ela mesma, formatada como uma base de dados. Em seguida, cada *Seed* é utilizada, por vez, em iterações sucessivas, tendo como banco de dados formatado as proteínas recrutadas pelo SeedServer. A quantidade de iterações, no entanto é limitada até que haja convergência ou pelo auto-escore, sendo a última iteração válida aquela onde o mesmo é superior a 70% do seu valor inicial. Esse procedimento ajuda a evitar um comportamento conhecido de deterioração da matriz após sucessivas iterações onde é possível observar incorporação de não-homólogos ou até mesmo em alguns casos a exclusão da própria sequência utilizada como *Seed*. O valor *default* de 70% foi estabelecido por observações em modelos de diversos grupos KO caracterizados no projeto de doutoramento em Bioinformática em andamento para anotação de proteínas hipotéticas com uso de PSI-BLAST do aluno Henrique de Assis Lopes Ribeiro pela UFMG, também responsável pelo desenvolvimento do módulo PSI-BLAST. Todas as sequências presentes nesse momento são consideradas PSI-validadas. A coluna referente à validação incorpora valor 1 para validadas, 0 para não validadas. Os valores de *E-value* são incorporados na tabela de resultados do Seed Linkage, sendo mostrado, em caso de múltiplas *Seeds*, o valor referente ao alinhamento de menor *E-value* (melhor alinhamento com uma *Seed*).

Por fim, é feita a comparação das estruturas secundárias através do já mencionado método SOV entre cada sequência recrutada e todas as *Seeds* presentes, sendo mostrado na tabela final o melhor valor obtido para cada sequência recrutada.

O processo completo será discutido melhor e apresentado na seção 6.1. Um breve exemplo do método de armazenamento dos resultados finais em uma tabela MySQL pode ser visto na Tabela 1.

Tabela 1: Exemplo de tabela final com resultados SeedServer. SL: Seed Linkage; KO: KEGG *Orthology*; TXID: identificador taxonômico numérico, SOV: *Secondary Structure Overlap* dado como porcentagem. UniProt_ac: identificador de seis caracteres do UniProtKB. Membro SL: 2 para *Seed*, 1 para recrutado-Seed Linkage e 0 para não recrutado-Seed Linkage. Membro KO: 1 para original do KO; 2 para pertencente ao UEKO e 0 para não pertencente ao KO ou UEKO.Tamanho: número de resíduos de aminoácidos. SwissProt/TrEMBL: 1 para SwissProt e 0 para TrEMBL. PSI: 1 para validado e 0 para não validado.

UniProt_ac	**Membro SL**	**Membro KO**	**Tamanho**	**TXID genoma**	**TXID filo**	**SwissProt/ TrEMBL**	**PSI**	**PSI *E-Value***	**SOV (%)**	**Descrição**
O26346	2	1	425	187420	28890	1	1	0	1	Histidina-tRNA ligase
B9AFG2	1	2	431	483214	28890	0	1	0	0.82	Histidina-tRNA ligase
F6D3U8	1	0	447	868131	28890	0	1	$2e^{-177}$	0.78	Histidina-tRNA ligase
P07178	0	2	508	10036	7711	1	1	$9e^{-155}$	0.73	Histidina-tRNA ligase
D8SXN2	1	0	935	88036	35493	0	0	$1e^{-4}$	0.67	Proteína não caracterizada

5.1.5. Relatório taxonômico final

Para cada agrupamento formado ao final de um processamento SeedServer, um relatório taxonômico é fornecido exibindo os diversos níveis taxonômicos com os dados apresentados no item 5.1.1 para cada proteína recrutada. Também através desses dados, para dar suporte a presença/ausência dos genes em um determinado clado, é possível determinar e agrupar a quantidade de genomas dentre os diferentes níveis de sequenciamento para qualquer clado existente. As sequências recrutadas são disponibilizadas em formato FASTA através do programa FASTACMD fornecido conjuntamente com o programa BLAST.

5.1.6. *Web Services*

Web Services (WS) são métodos conhecidos para comunicação entre programas, ou seja, uma vez programados, a intervenção humana é desnecessária para que haja o processamento dos dados. Como a comunicação entre as partes é feita através de uma linguagem padronizada chamada XML (eX*tensible Markup Language*) oferece a vantagem da interação entre aplicações que podem ser desenvolvidas em diferentes linguagens de programação. Os serviços ou métodos oferecidos por um WS bem como parâmetros e formato dos resultados obtidos são descritos em um documento WSDL (*Web Services Description Language*).

Foram desenvolvidas em nosso Laboratório de Biodados na UFMG, como parte do projeto de doutoramento em Bioinformática do aluno Henrique Velloso, diferentes WSs. No presente trabalho foram utilizados dois tipos: (i) - *Bioinfomatics Open Web Services* (BOWS): ferramenta construída com o objetivo de integrar em um único banco de dados diferentes aplicações bioinformáticas que podem ser requisitadas por meio de WS e (ii) – *Lowest Common Ancestor Web Service*

(LCAWS): ferramenta capaz de encontrar o ancestral comum mais recente (LCA) dado um conjunto de identificadores taxonômicos de diferentes organismos. O LCAWS trabalha com 18 níveis taxonômicos tradicionais da classificação de Lineu (e. g.: reino, classe, ordem, subfamília, espécie e organismo seqüenciado) além dos níveis intermediários existentes na base *Taxonomy* do NCBI que não possuem classificação definida (e. g.: na taxonomia humana, o grupo *Vertebrata* está abaixo do subfilo *Craniata*, porém é definido como *no rank* ou sem classificação).

Nesse sentido, como forma de otimizar o processamento, a ferramenta SeedServer foi cadastrada no banco de dados BOWS, além dela mesma requisitar o serviço LCAWS para os agrupamentos de homólogos formados pelo SeedServer, sendo o uso dessas funcionalidades transparentes ao usuário comum (Figura 4).

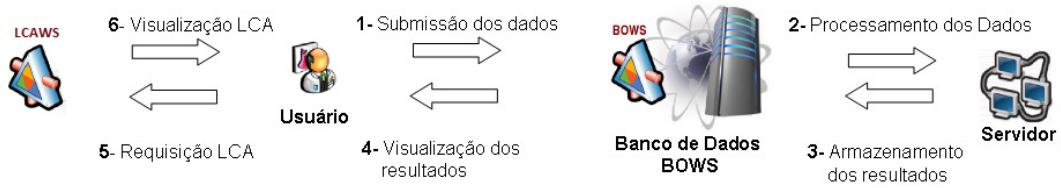


Figura 4: Esquema ilustrando a utilização dos *Web Services* BOWS e LCAWS através do SeedServer.

5.1.7. Página *web* SeedServer

Para disponibilização da ferramenta SeedServer na *web* foi criada uma página em HTML (*HyperText Markup Language*) tendo o método CGI (*Common Gateway Interface*) como interface entre os dados dos usuários e o programa SeedServer.

5.2. Metodologia para estudos de aminoácidos essenciais

Como estudo de caso para utilização da ferramenta SeedServer, foram analisados agrupamentos de homólogos para proteínas das vias biossintéticas de aminoácidos essenciais.

5.2.1. Bancos de Dados e programas

Os bancos de dados, bem como os programas utilizados para criação de grupos de homólogos e análises filogenéticas, estão descritos no item 5.1. Para estas análises, em casos indicados na seção de Resultados, as sequências UniProtKB descritas como fragmentos foram incluídas.

5.2.2. Procedimento de inspeção das vias

As vias biossintéticas dos AEs foram manualmente inspecionadas utilizando KEGG *Pathway* [39]. Identificadores UniProtKB de organismos modelo, conhecidamente autotróficos, *Saccharomyces cerevisiae*, *Arabidopsis thaliana e Pyrococcus horikoshii* (incluído para o estudo da via alternativa de lisina) foram obtidos para as respectivas enzimas e utilizados como *Seeds* na ferramenta SeedServer.

Adicionalmente, realizamos análises com BLASTp (limiar de *E-value* de 10^{-10}) utilizando como *queries* todas as sequências agrupadas pelo SeedServer e como base de dados o UniProtKB (incluindo sequências fragmentadas) numa tentativa de capturar (i) CDSs (*Coding* DNA *sequence*) parciais resultantes de sequenciamentos incompletos de mRNA e (ii) proteínas modeladas incorretamente a partir do genoma, onde a metionina inicial não foi identificada. Para evitar alinhamentos

parciais entre domínios protéicos que não refletem a funcionalidade da proteína como um todo, um filtro adicional exigindo 50% de identidade e 50% de cobertura da *query* pelo alinhamento BLAST foi aplicado. Os resultados dessas buscas são representados com triângulos na seção de Resultados.

5.2.3. Análises de distâncias filogenéticas

Realizadas como descrito no item 5.1.2. As distâncias entre os grupos taxonômicos foram obtidas das árvores, fornecidas pela ferramenta MEGA5, tendo ancestrais de *Streptophyta*, *Dikarya* e grupos de metazoários como referência. Somente ramos com suporte de *bootstrap* acima de 50 foram considerados. A razão utilizada foi calculada da seguinte forma:

Distância D-P / Distância S-D

onde D (com sentido direcional "de") é o ramo que agrupa sequências de animais metazoários até P (com sentido direcional "para"), referente ao ramo ancestral de *Streptophyta* ou então de *Dikarya*; A distancia S-D é a encontrada entre *Streptophyta* e *Dikarya*, respectivamente. Assim, a razão acima tem no denominador a distância entre as sequências de planta e fungo. Números sensivelmente maiores que 1 indicam portanto que as distâncias do ramo animal para os ramos fungo ou planta, são sensivelmente maiores que a distância entre os ramos fungo e planta.

5.3. Propagação dos termos de ontologia funcional protéica

A relação completa de todos os termos de ontologia funcional associados aos identificadores UniProtKB foram obtidos do projeto GOA (*Gene Ontology Annotation*) em: [ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/UNIPROT/]. Foi

utilizada a versão disponível em 27 de novembro de 2012. Para inferir correlação e ordem na hierarquia foi utilizado o arquivo:

[http://www.geneontology.org/ontology/obo_format_1_2/gene_ontology.1_2.obo]

Os termos associados às sequências protéicas provenientes de todos os organismos das 82 espécies [http://www.pantherdb.org/panther/summaryStats.jsp] presentes no banco de dados PANTHER foram escolhidos como ponto de origem da propagação uma vez que contam com curadoria manual.

Para subsidiar a propagação foi feito um mapeamento completo da relação de cobertura entre todos os membros de grupos UniRef50 contendo duas ou mais sequências de pelo menos um dos organismos citados acima, sendo a ferramenta resultante desse processo nomeada U-MAGE (*UniRef50 Matrices for Annotation of Gene Ontology Entries*).

A cobertura foi calculada da seguinte forma: Tamanho do alinhamento obtido por BLAST dividido pelo tamanho da proteína. Como o denominador varia de acordo com a proteína para cada par de alinhamento, matrizes foram escolhidas como forma de armazenamento dos dados (Tabela 2). Para cada sequência representante de uma linha é possível selecionar a partir de um dado valor de cobertura limite outras sequências que poderão receber sua anotação funcional, consequentemente, cada representante em uma coluna recebe anotação de outras sequências presentes em suas linhas.

Por exemplo, considerando-se um limite de cobertura de 80%, Q9BYK8 representado na Tabela 2, poderia doar sua anotação às sequências A2AS03, B9A0U1 e D3ZFS4, mas não para Q7TT03. Em contrapartida, Q7TT03 poderia doar seus termos à Q9BYK8, além de A2AS03, B9A0U1 e D3ZFS4.

Tabela 2: Matriz hipotética contendo dados de cobertura de membros UniProtKB de mesmo UniRef50. Valores da diagonal representados por um zero. Valores equivalentes a cem representados por dois zeros.

	Q9BYK8	**A2AS03**	**B9A0U1**	**Q7TT03**	**D3ZFS4**
Q9BYK8	0	00	00	9	00
A2AS03	91	0	99	7	00
B9A0U1	92	00	0	8	00
Q7TT03	00	95	00	0	95
D3ZFS4	90	00	99	7	0

Sentidos da propagação tendo Q9BYK8 como referência: propagação "para" sequências destacas em azul e "de" sequências destacadas em laranja, considerando cobertura de 90%.

A cada anotação está associada um Código de Evidência, de duas ou três letras, que dá suporte e qualifica a informação anotada, podendo ser por critérios manuais como Inferido por Ensaio Direto (IDA – *Inferred from Direct Assay*) ou automáticos, como Inferido por Alinhamento de Sequência (ISA - *Inferred from Sequence Alignment*). Uma lista completa dos Códigos de Evidencia está disponível em [http://www.geneontology.org/GO.evidence.shtml]. Termos propagados pela metodologia U-MAGE são atribuídos com novos códigos de evidência, caracterizados pela primeira letra 'U' (e. g.: UDA e USA).

Termos marcados com o Qualificador (*Qualifier* = *'NOT'*), o que significa que o produto gênico não esta associado ao termo GO não são considerados. Os termos GO:0005515 (ligação a proteína) e GO:0003674 (Função molecular) também são desconsiderados por serem termos genéricos. Somente relações do tipo "is_a" foram representadas.

As matrizes foram calculadas com auxílio do super-computador veredas do Centro Nacional de Processamento de Alto Desempenho (CENAPAD) da UFMG com a versão UniRef50_2012_08, produzida com o novo filtro de cobertura de 80%,

adquirida diretamente com membros mantenedores do UniRef50, porém não disponível publicamente.

A estrutura geral da organização dos dados em tabelas locais MySQL está representada na Figura 5.

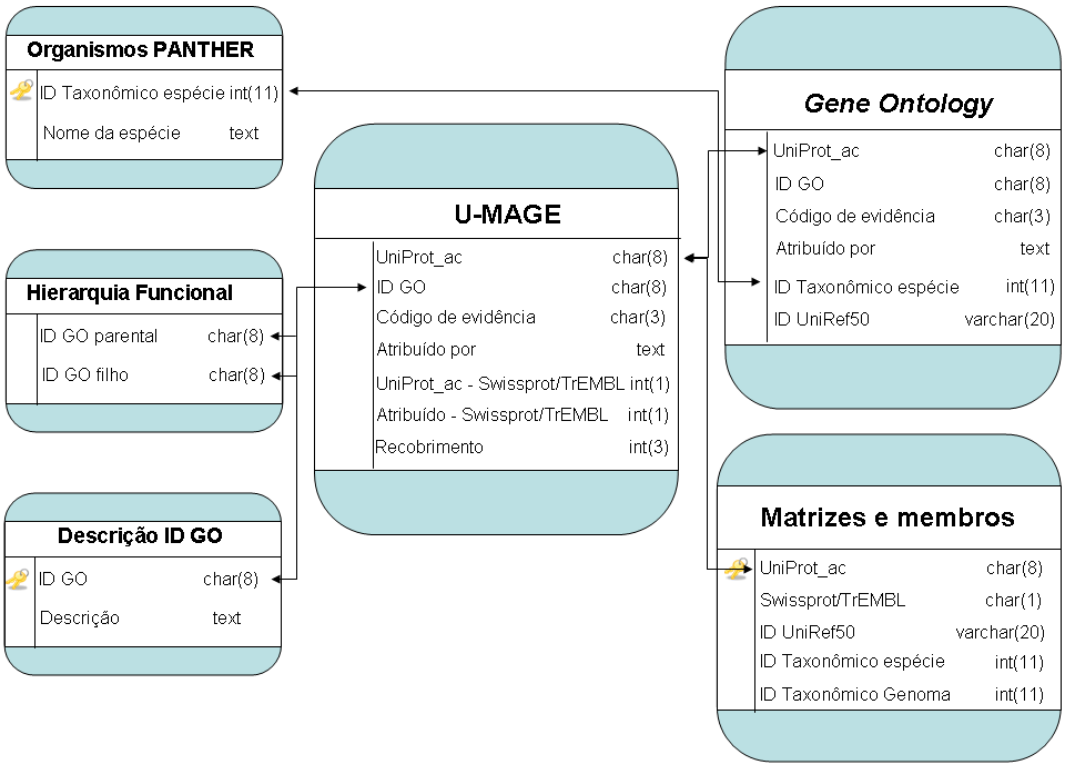


Figura 5: Mapa físico do banco contendo as tabelas utilizadas pelo U-MAGE. As setas pretas indicam os identificadores que correlacionam as tabelas. As chaves indicam quais são os identificadores únicos. ID: identificador; GO: *Gene Ontology*.

5.3.1. Página *web* U-MAGE

Desenvolvida com descrito no item 5.1.7.

6. Resultados

6.1.SeedServer

Uma meta deste trabalho foi a integração de abordagens utilizadas em nosso grupo de pesquisa para propagar a informação associada a uma proteína para um grupo de proteínas relacionadas à mesma. Assim foi concebido o sistema SeedServer. Uma de suas bases é o programa Seed Linkage [15], que realiza buscas de ortólogos e parálogos com base em buscas do tipo BBH (*Best Bidirectional Hit*, ou "melhor alinhamento recíproco"). Outra fonte de informação muito utilizada no grupo é uma base de dados enriquecida a partir da base de dados KEGG *Orthology*, que é produzida por uma metodologia muito similar à que fora utilizada para o enriquecimento da base COG [16]. O enriquecimento da base COG foi denominado UECOG (UniRef50 *Enriched* COG) e, analogamente, o enriquecimento de KO é chamado UEKO. O procedimento de enriquecimento consiste em recrutar membros de grupos UniRef50, produzidos pelo consórcio UniProt, para grupos da base a ser enriquecida, desde que eles contenham algum membro em um grupo UniRef50. Na produção de UECOG, somente aqueles candidatos que tivessem no máximo 10% de diferença de tamanho com o recrutador eram recrutados. Basicamente no procedimento atual para geração do UEKO, recrutadores são alinhados localmente por intermédio de BLAST aos candidatos e somente aqueles candidatos cujo alinhamento apresenta mais que 50% de cobertura do recrutador são recrutados. A base UEKO continua sendo atualizada por nosso grupo a partir da última versão pública do KO e utilizando os grupos UniRef50 mais atuais. Para uma melhor utilização do programa Seed Linkage é ideal que as buscas sejam feitas em bases de dados contendo apenas proteínas completas. Assim, um simples filtro da base de

dados UniProt, eliminando-se todas que contenham a marcação *fragment* na linha de descrição, gera uma base de dados contendo apenas proteínas completas. Isso é interessante porque na execução do Seed Linkage e em alguns passos do Sistema SeedServer é utilizado o parâmetro cobertura, que é mais fidedigno quando somente proteínas completas estão sendo comparadas. A execução do sistema SeedServer requer, portanto, a escolha de um identificador UniProtKB de uma sequência classificada como *Complete*, o que é facilmente verificável na página da proteína. No caso da proteína de interesse não estar presente na base UniProtKB, recomenda-se a utilização do identificador de uma proteína bastante similar, o que pode ser obtido com a execução de BLAST no site do UniProt [http://www.uniprot.org/].

Como descrito na seção Metodologia 5.1.3, o sistema SeedServer inicia o agrupamento executando o programa Seed Linkage. Este programa cria grupos a partir de cada sequência fornecida como semente (*Seed*), bem como a partir de seus possíveis parálogos, os quais são recrutados pelo programa. Quando mais de uma *Seed* é utilizada, o programa Seed Linkage termina por juntar os grupos que tenham mais que 50% de membros em comum. Assim sendo, Seed Linkage pode formar um único grupo, ou diversos e isso é mantido na execução do sistema SeedServer. É comum usuários submeterem proteínas que suspeitam serem homólogas como *Seeds* e o programa Seed Linkage irá reuni-las ou criar grupos distintos, já informando ao usuário se as *Seeds* formam um ramo evolutivo ou múltiplos, caso sejam similares, já que observa-se em inspeção manual uma correlação entre ramos e grupos Seed Linkage distintos. Subsequentemente, como mostrado na Figura 6, o sistema verifica se a *Seed* é encontrada em grupos KO e grupos UEKO e recruta os membros complementares desses grupos, em um processo rápido já que se trata de consulta a banco de dados. Na página de resultados é mostrado se as sequências

recrutadas pertencem ao KO, se pertencem exclusivamente à porção que o enriquece (UEKO mas não KO), ou se a nenhuma delas.

Pode ocorrer de a sequência *Seed* não ocorrer nas bases de dados KO e UEKO. Neste caso os homólogos seriam agrupados somente pelo programa Seed Linkage. Todavia, quando isso ocorre, o mesmo procedimento utilizado para gerar o UEKO é utilizado com a *Seed*. Por analogia, esse passo é chamado UE-*Seed* e consiste em identificar o grupo UniRef50 ao qual a *Seed* pertence, alinhá-la aos membros deste grupo e recrutar sequências cujo alinhamento apresenta cobertura maior ou igual a 50% da *Seed*.

O sistema SeedServer utiliza informações taxonômicas frequentemente e é plausível que usuários queiram evitar trabalho adicional de filtrar genes que não pertençam a clados de interesse, assim sendo, somente as sequências recrutadas pertencentes ao clado escolhido pelo usuário são consideradas. Isto reduz o custo computacional da execução do Seed Linkage.

Como mostrado na Figura 6 e detalhado na seção Metodologia 5.1.4, segue-se ao recrutamento uma validação dos agrupamentos com PSI-BLAST. No procedimento, todas as sequências então recrutadas compõem uma base de dados que é formatada e as sequências *Seed* são utilizadas como entrada do PSI-BLAST. Com base em um projeto em desenvolvimento no grupo (H. A. L. Ribeiro e colaboradores, não publicado) é imposto um limite de auto-escore à *Seed* na execução do PSI-BLAST. O *default* é 0,7. Isto significa que o PSI-BLAST reporta a última iteração onde o auto-escore (o escore da *Seed* contra ela mesma) é superior a 70%. Como as matrizes PSSM (*Position-specifc scoring matrix*) são atualizadas a cada iteração, é comum a *query* do PSI-BLAST obter escores cada vez mais baixos quando as iterações começam a capturar proteínas divergentes. Desta maneira, a busca é

interrompida antes que a matriz se deteriore. Para que uma sequência recrutada permaneça após essa validação, ela deve apresentar um valor de *E-value* abaixo do *cutoff* para alguma das *Seed*s.

Como o programa Seed Linkage utiliza relações de BBH formadas pelo programa BLASTp, algumas sequências com *E-value* próximo do limite são descartadas neste ponto. Além disso, quando alguma sequência é agrupada pelo KO por causa de um pequeno domínio em comum com o restante do grupo, ela e seus relativos no UEKO são eliminadas do agrupamento, pois o PSI-BLAST tem a *Seeds* como sequências *query*.

Neste momento os grupos já estão formados, todavia o sistema ainda executa uma comparação da possível sobreposição de estrutura secundária entre *Seeds* e recrutadas, executando a predição de estrutura secundária com o programa Predator [32], (o qual pretendemos substituir pelo programa SSPro4 [40] por ser mais preciso quando o sistema estiver disponível em melhores servidores) e comparando as estruturas secundárias preditas calculando o parâmetro SOV [34], o qual varia entre 0 (totalmente distintas) e 1 (idênticas).

Em resumo, para criação e validação de grupos de sequências protéicas homólogas a partir de uma ou mais sequências de interesse, foram agregados diferentes bancos de dados e programas, coordenados por um programa principal escrito na linguagem de programação PERL, resultando em uma ferramenta denominada SeedServer (Figura 6).

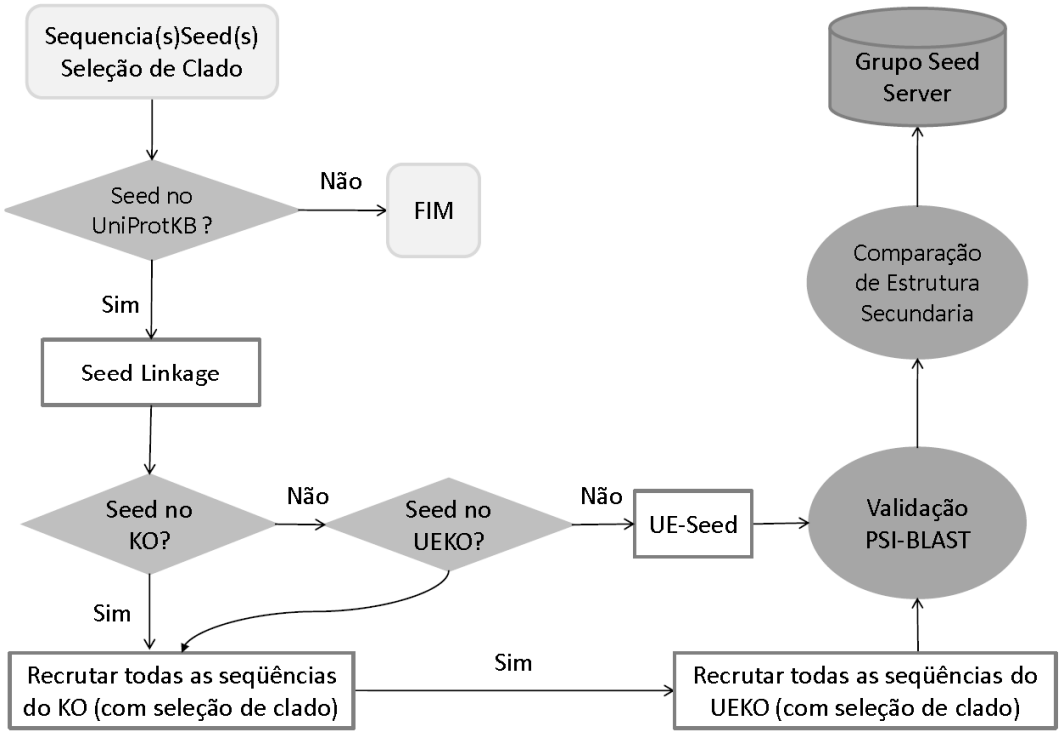


Figura 6: Fluxograma com procedimento SeedServer completo. KO: KEGG *Orthology*; UEKO: *UniRef50 Enriched* KO; UE-Seed: UniRef50 *Enriched-Seed.* A 'seleção de clado' consiste na filtragem de táxons não pertencentes à taxonomia escolhida pelo usuário.

6.1.1. Aminoacil-tRNA sintetases

6.1.1.1. Agrupamentos

Para exemplificar e testar o uso da ferramenta SeedServer foram realizados agrupamentos de aminoacil-tRNA sintetases (aaRS) utilizando como *Seeds* todas as sequências de archaeas presentes nos respectivos agrupamentos do banco de dados COG (*Clusters of Orthologous Groups*), sendo o universo de busca todas as outras proteínas completas disponíveis no UniProtKB. Uma aaRS é uma enzima capaz de catalisar a esterificação de um aminoácido específico, ou de seu precursor, ao seu tRNA correspondente. Essas aaRSs foram escolhidas devido a sua ubiqüidade,

proporcionando um conjunto rico de testes dentre todos os reinos da vida. A escolha de sequências de archaea como *Seed* nos permite testar a capacidade de recrutamento de homólogos em outros clados. O uso de sequências do COG, limitado a poucos genomas em comparação aos disponíveis atualmente, simula a necessidade de um usuário em propagar informação a partir de um conjunto específico de *Seeds*.

Considerando que nosso procedimento de buscas envolve os grupos KO e, para minimizar as variáveis envolvidas, somente aaRSs associadas com uma única entidade KO foram escolhidas, totalizando 13 experimentos (Tabela 3). Além disso, é relatado que algumas aaRSs possuem mais que uma atividade [41] e envolvê-las prejudicaria a conferência pelo número de catálogo enzimático (EC – *Enzyme Comission*), descrita posteriormente.

Para nove experimentos, o programa Seed Linkage formou grupos únicos e dividiu os quatro restantes em sub-grupos (Tabela 3), embora as sequências dos subgrupos estejam reunidas na base KO. Em suporte ao resultado apresentado pelo Seed Linkage, a precisão desses sub-agrupamentos de homólogos mais próximos pode ser conferida na distribuição das *Seeds* em árvores filogenéticas (Figura 7).

O esquema no formato de Diagrama de Venn mostrado na Figura 8 representa todas as categorias possíveis a que uma sequência possa pertencer ao final de um procedimento SeedServer, enquanto a Tabela 3 resume os resultados obtidos para as 13 aaRSs.

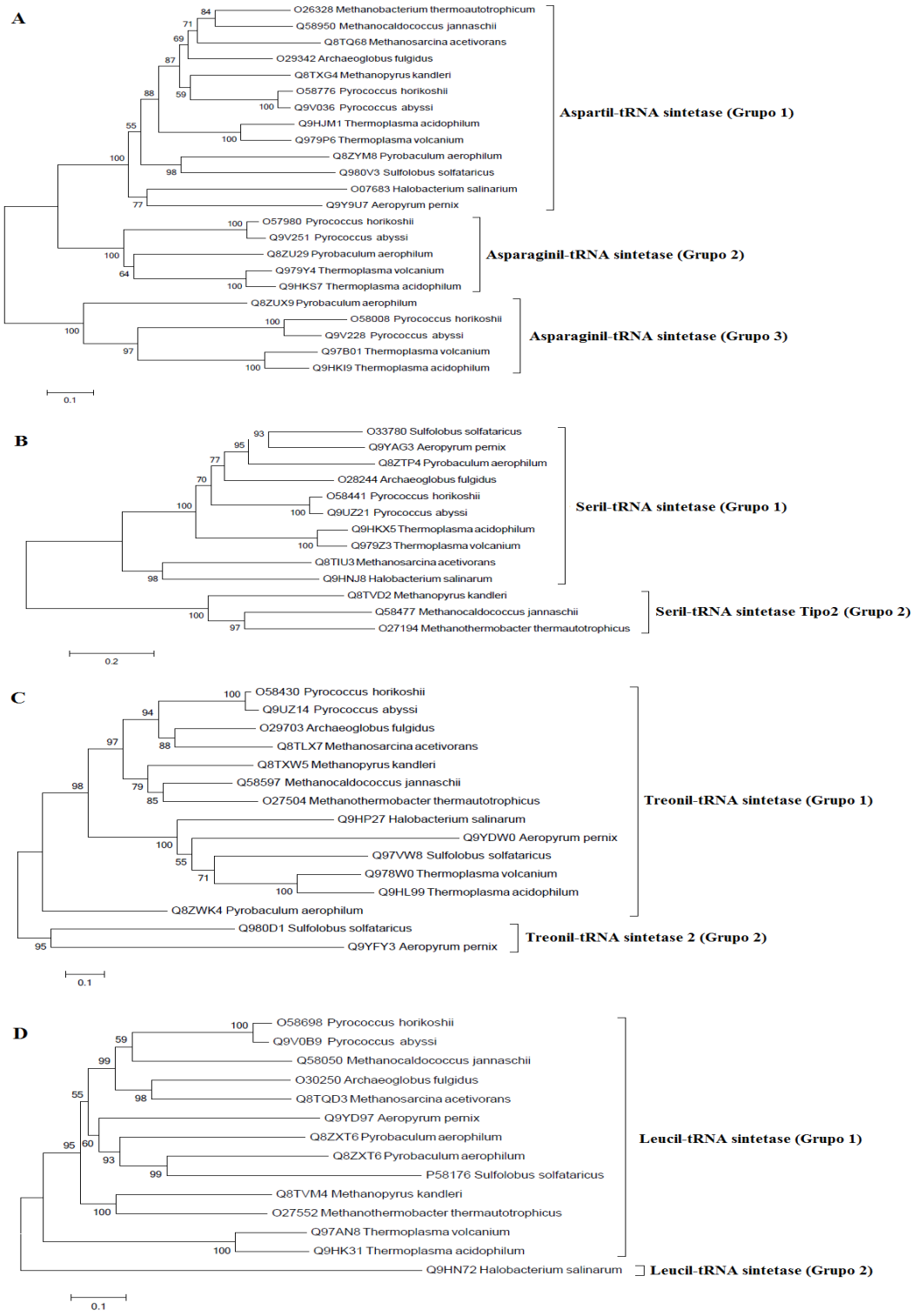


Figura 7: Árvores filogenéticas contendo *Seeds* de archaeas para sub-grupos de aminoacil-tRNA sintetases divididos pelo Seed Linkage. A: Aspartil/Asparaginil-aaRS; B: Seril-aaRS; C: Treonil-aaRS; D: Leucil-aaRS.

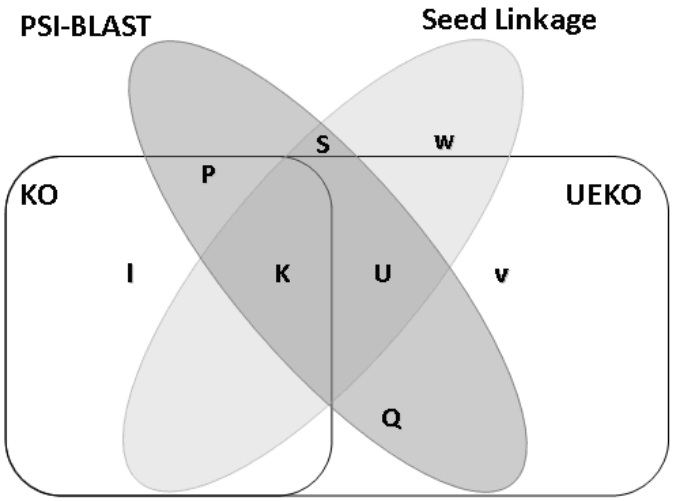


Figura 8: Diagrama de Venn representando as categorias existentes ao final do agrupamento SeedServer. S: Sequências agrupadas pelo Seed Linkage, mas não presentes no KO ou UEKO e PSI-validadas; K: Agrupadas pelo Seed Linkage, presentes no KO e PSI-validadas; U: Agrupadas pelo Seed Linkage, presentes no UEKO e PSI-validadas; P: Presentes no KO e PSI-validadas; Q: Presentes no UEKO e PSI-validadas; l: Presentes no KO e não PSI-validadas; v: Presentes no UEKO e não PSI-validadas; w: Agrupadas pelo Seed Linkage somente e não PSI-validadas. KO: KEGG *Orthology*; UEKO: UniRef50 *Enriched* KO.

As categorias representadas pelas letras maiúsculas S, K, U, P e Q contém apenas sequências PSI-validadas. A categoria S agrega proteínas recrutadas somente pelo Seed Linkage, ausentes tanto no KO quanto UEKO. Observamos que 26% e 46% das proteínas nessa classe são derivadas de genomas incompletos/sem projeto ou em processo de montagem, respectivamente. Isso ressalta importância da utilização do Seed Linkage, um programa planejado para buscas de homólogos em organismos sem genoma completo. Quase 2% do recrutamento total foi feito com Seed Linkage e foi PSI-validado. Essa proporção varia dependendo do tipo de *Seed* utilizada. Neste caso, como se trata de uma família de enzimas muito bem caracterizada, a contribuição de 2% é relevante por completar o universo de busca, mas em várias situações as proteínas estão ausentes no UEKO.

A categoria K agrega proteínas recrutadas pelo Seed Linkage e presentes em algum grupo KO, enquanto a U representa sequências agrupadas pelo Seed Linkage e pertencentes ao UEKO somente (Tabela 3). A contribuição do KO, de cerca de 10%, foi elevada neste caso, o que é esperado dado que se trata de enzimas prontamente anotáveis em genomas inclusos na referida base. Nota-se a também elevada contribuição do UEKO, aproximadamente 16%, explicada pelo fato do enriquecimento recrutar sequências de genomas não completos, fora de KO.

Algumas vezes o programa Seed Linkage não detecta homólogos mais distantes às *Seeds* e membros do KO (categoria P) ou UEKO somente (categoria Q), no entanto elas são agrupadas e validadas. Isso pode ser explicado pelo fato do Seed Linkage ser um método de agrupamento mais rigoroso quanto à similaridade de sequências e é um resultado esperado, já que o recrutamento envolve similaridade e cobertura superiores a 50% em seu modo *default*.

Considerando os dados das classes S, K e U mostrados na Tabela 3, fica clara a importância na participação do Seed Linkage com milhares de recrutamentos rigorosos, representando 28,13% do total, mesmo quando os experimentos são iniciados com um número limitado de *Seeds* (até 15 nesse experimento) e especificamente de archaeas. Adicionalmente, considerando os dados das categorias U e Q, fica evidente o número de sequências provenientes do UEKO (63,27%). Considerando-se que KO é descarregado e UEKO é pré-computado, a busca adiciona muita informação com baixo custo computacional.

Levando-se em consideração somente PSI-validadas, 40%, 46% e 14% das proteínas são, respectivamente, derivadas de genomas completos, em processo de montagem e incompletos/sem projetos.

Por outro lado, classes representadas pelas letras minúsculas l, v e w contém sequências rejeitadas pelo processo de validação PSI-BLAST. A categoria l é surpreendente, uma vez que está relacionada ao agrupamento KO. A categoria v é similar à anterior, mas aqui o PSI-BLAST rejeita sequências da porção exclusiva do UEKO e, finalmente, a categoria w reúne rejeições de sequências encontradas pelo Seed Linkage somente. Essas rejeições operacionalmente constituem sequências que não obtiveram um *E-value* abaixo do *cutoff* quando alinhadas com quaisquer *Seeds*. Esses eventos não chegam sequer a 0,05% quando somados.

Tabela 3: Resultados do agrupamento SeedServer para as aminoacil-tRNA sintetases.

COG/KO	aaRS/EC	*Seeds*	S	K	U	P	Q	l	v	w	Total
COG0013 / K01872	Ala (6.1.1.7)	13	64	600	1007	877	1599	0	0	0	4160
COG0017 / K01876	Asp (6.1.1.12)	13	287	199	170	1345	2701	6	2	0	4723
COG0017* / K01893	Asn (6.1.1.22)	10	7	203	391	731	1522	0	0	0	2864
COG0018 / K01887	Arg (6.1.1.19)	13	85	557	1165	943	1554	0	0	0	4317
COG0060 / K01870	Ile (6.1.1.5)	13	24	268	401	1264	2405	0	0	0	4375
COG0124 / K01892	His (6.1.1.21)	14	104	626	1194	867	1535	0	0	2	4341
COG0143 / K01874	Met (6.1.1.10)	13	54	513	905	1075	1992	0	0	0	4552
COG0162 / K01866	Tyr (6.1.1.1)	14	91	306	290	1271	2743	1	4	1	4721
COG0172* / K01875	Ser (6.1.1.11)	14	61	477	1024	1074	1906	0	0	0	4555
COG0180 / K01867	Trp (6.1.1.2)	14	65	252	252	1364	2888	0	6	0	4871
COG0441* / K01868	Thr (6.1.1.3)	15	51	697	1320	871	1531	0	0	0	4484
COG0495* / K01869	Leu (6.1.1.4)	14	93	433	415	1104	2303	0	0	0	4362
COG0525 / K01873	Val (6.1.1.9)	13	40	398	637	1074	2037	0	0	0	4199
Total	-	173 (0,3%)	1026 (1,81%)	5529 (9,78%)	9171 (16,22%)	13860 (24,52%)	26716 (47,26%)	7 (0,01%)	12 (0,02%)	2 (0,00%)	56524

Distribuição nas diferentes categorias das sequências recrutadas pelo SeedServer para as 13 aminoacil-tRNA sintetases (aaRS). COG: *Clusters of Orthologous Groups*; KO: KEGG *Orthology*; EC: *Enzyme Comission*. Descrição das categorias S, K, U, P, Q, l, v e w assim como indicado na Figura 8. *COGs divididos pelo Seed Linkage.

6.1.1.2. Casos de rejeição na validação PSI-BLAST

- Para aspartil-tRNA sintetase, seis proteínas na categoria 1 foram consideradas inválidas pelo método de validação PSI-BLAST. Q8BJY7, C4QJ96, Q6MTH6 e E4PUG3, com respectivamente 87, 27, 100 e 90 resíduos de aminoácidos, não atingiram valores satisfatórios de corte. A média de tamanho das *Seeds* é 430 aminoácidos. E2LXX2 e Q5FNR4, embora com, respectivamente, 431 e 311 resíduos, também foram rejeitas, mas esse resultado é suportado por valores de identidade inferiores a 39% com outras sequências do grupo e nenhuma família PANTHER foi atribuída a elas pela ferramenta PANTHER HMM *scoring tool* (ver métodos e Tabela 6 adiante). D9QXB3 e F8K681 da categoria v, com 100 e 90 resíduos respectivamente, também foram rejeitadas, já que baixas identidades em adição aos curtos alinhamentos colaboram para que não alcancem o limiar de corte.
- Para histidil-tRNA sintetase, Seed Linkage encontrou duas proteínas que foram posteriormente rejeitadas (D8SPU1, D8SXN2) e uma (D8S8H4) PSI-validada e presente no UEKO, todas da planta *Selaginella moellendorffii*. D8S8H4 contem 1.665 resíduos, mas somente a porção N-terminal (1086 até 1504 aproximadamente) é relacionada à função de histidil-tRNA sintetase, enquanto a outra porção esta relacionada à atividade RNA helicase (Tabela 6). Esse recrutamento correto permitiu que o Seed Linkage recrutasse durante a busca de parálogos em *S. moellendorffii,* os outros dois falsos positivos mencionados anteriormente com suporte de cobertura, mas com função única de RNA helicase, mas que foram por fim corrigidos pela validação.

- Para tirosil-tRNA sintetase, Seed Linkage recrutou a sequência B3N840 de 320 resíduos que não foi validada. Com o aumento dos seus parâmetros de identidade e cobertura de 50% para 60% esse recrutamento não é observado. Como se trata de um recrutamento próximo ao limiar, a troca da matriz de substituição utilizada pelo BLASTp pela matriz PSSM gerada pelo grupo formado no processo de validação PSI-BLAST evitou este recrutamento. As outras rejeições foram uma do KO (F7UM56) e quatro da porção enriquecida do UEKO (H0P2J8, H0P6A6, H0PKD2 e P73144), sendo todas com 78 resíduos de aminoácidos apenas, pequenas se comparadas com a média de 340 resíduos das *Seeds*, sendo improvável alcançarem o limiar de *E-value* pela concorrência de alinhamento pequeno e baixa identidade.
- Para triptofanil-tRNA sintetase, outro caso de proteínas multi-domínios foi observado. B8NCZ6 presente no K01876 possui funções de proteína ribossomal e triptofanil-tRNA sintetase. Isso permitiu o recrutamento de Q5AYZ1 da porção exclusiva do UEKO, que é uma proteína ribossomal somente. Esse agrupamento errôneo também foi detectado e corrigido pelo método PSI-BLAST.

6.1.1.3. Correlação entre número EC e validação PSI-BLAST

A curadoria manual de algumas proteínas rejeitadas pelo método de validação PSI-BLAST mostrou casos de verdadeiros negativos, no entanto essa abordagem seria inalcançável na inspeção de positivos (as proteínas validadas), uma vez que se trata de milhares de sequências.

Nesse sentido, para investigar os casos positivos, foi obtido o número EC associado às *Seeds* e a todas as sequências recrutadas, sempre que tal associação estava disponível em dois experimentos distintos, nomeados A e B. No experimento A (Tabela 4) as *Seeds* de aminoacil-tRNA sintetases foram comparadas contra o banco de dados protéico completo e sequências SwissProt analisadas. A Tabela 4, coluna 4 mostra que, considerando somente proteínas manualmente curadas, onde um maior nível de segurança pode ser atribuído ao número EC em questão, 100% das sequências agrupadas e validadas pelo SeedServer, apresentaram o número EC esperado em nove dos treze casos (Ala, Asn, Arg, Ile, His, Met, Tyr, Trp e Thr). A análise global dentre entradas SwissProt resultou em 99,7% de verdadeiros positivos indicando alta precisão. Para os quatro casos restantes foram observados os respectivos cenários:

- Aspartil-tRNA sintetase (Asp) EC:6.1.1.12 : De um total de 748 proteínas, oito (A4J412, A9WA97, B8GBG7, B9LCU8, Q5WGB1, Q65I63, Q6G9A8 e Q8NWP3) de dois filos bacterianos (*Chloroflexi* e *Firmicutes*) foram recrutadas exclusivamente pelo Seed Linkage mas estão descritas como Asparaginil-tRNA sitetase (EC: 6.1.1.22). No entanto as comparações de suas estruturas secundárias com as *Seeds* de archaea mostram correlações entre 73% e 80%, além de todas elas terem sido atribuídas à família PANTHER PTHR:22594, subfamília SF6, correspondente a aspartil-tRNA sintetase, em suporte ao recrutamento feito pelo Seed Linkage. Isso indica que o provável erro deve-se a alta similaridade existente entre os dois tipos de aaRSs.
- Seril-tRNA sintetase (Ser) EC:6.1.1.11: De um total de 817 proteínas, duas (Q89GR3 e Q89VT8) da Proteobacteria *Bradyrhizobium japonicum*

foram recrutadas exclusivamente pelo Seed Linkage mas estão descritas como Aminoacil-[proteína acil-carreadora] sintetase EC:6.2.1.n2. Sobreposição de respectivamente 77% e 72% de estrutura secundária foi observada com as *Seeds*, no entanto nenhuma família ou subfamília PANTHER foram atribuídas a essas duas sequências. Todavia, quando ambas sequências são utilizadas como *Seed* no SeedServer, Q89VT8 encontra Q8TVD2, entrada SwissProt, EC:6.1.1.11, com *E-value* de 10^{-80}; Q89GR3 não encontra nenhuma entrada SwissProt mas alinha com E5AVF5, Seryl-tRNA synthetase (EC 6.1.1.11), com *E-value* de $3x10^{-136}$, o que deixa a questão de serem realmente falsos positivos em aberto.

- Leucil-tRNA sintetase (Leu) EC:6.1.1.4: De um total de 764 proteínas, três (A8F8Q3, B9K8X3 e Q9X2D7) do gênero bacteriano *Thermotoga* foram recrutadas exclusivamente pelo Seed Linkage mas estão descritas como Valil (EC:6.1.1.9) ou Isoleucil-tRNA sintetases (EC:6.1.1.5). Similaridades estruturais de respectivamente 76%, 75%, 72% foram observadas. A alta similaridade entre esses três tipos de aminoacil-tRNA sintetases provavelmente se dá por terem divergido mais recentemente na evolução [42]. A mesma família PANTHER PTHR11946 (Iso, Leu, Tyr, Val e Met-tRNA sintetases) foi atribuída a essas sintetases em suporte ao resultado encontrado pelo Seed Linkage.
- Valil-tRNA sintetase (Val) EC:6.1.1.9: De um total de 343 proteínas, oito (A4XKR2, A7HMY6, B7IGT1, B8CWL4, B9K8X3, B9MRZ5, Q2RK59 e Q8R9L3) dos filos bacterianos *Firmicutes* e *Thermotogae* foram recrutadas exclusivamente pelo Seed Linkage, mas estão descritas como Isoleucil-tRNA sintetase EC: 6.1.1.5. Novamente uma alta similaridade

de estruturas secundárias foi observada variando de 72% a 76%, mesmo entre reinos diferentes. Novamente a família PANTHER PTHR11946 (Iso, Leu, Tyr, Val e Met-tRNA sintetases) foi atribuída, mostrando o mesmo quadro de agrupamento entre essas aminoacil-tRNA sintetases, altamente similares.

Alguns poucos casos de trocas entre essas sintetases também são observados entre proteínas TrEMBL.

O experimento B foi realizado de forma controlada, onde as *Seeds* de aaRSs foram comparadas com um banco de dados contendo somente as treze sintetases estudadas. Isso possibilitou a obtenção dos valores de verdadeiros e falsos positivos e negativos e consequentemente o cálculo dos coeficientes de i- *F-Score* e ii- MCC (*Matthews correlation coefficient*) que refletem o desempenho do teste realizado. O índice *F-Score* varia entre 0 (acurácia mínima) e 1 (acurácia máxima) enquanto o de MCC de -1 (acurácia mínima) a 1 (acurácia máxima). Similarmente, uma alta correlação entre o número EC de sequências PSI-validadas e o número EC correto também foi observada incluindo-se na análise sequências não manualmente curadas TrEMBL (Tabela 4, coluna 10, precisão). As taxas de falso negativo (sequências com o EC correto não detectadas) e falso positivo (sequências detectadas com o EC errado) mostraram-se baixas, como observado respectivamente nas colunas 7 e 9 da Tabela 4. Considerando todo o universo de proteínas presentes na base de dados UniProtKB (Tabela 4, décima primeira coluna), a revocação dos números EC relacionados às aminoacil-tRNA sintetases em estudo foi superior a 90% em todos os treze casos. Os valores de *F-Score* obtidos para os treze experimentos foram próximos a 1 em todos os casos, assim como os de MCC, indicando uma alta correlação entre a validação baseada em PSI-BLAST e anotação de número EC.

Tabela 4: Correlação entre número EC e validação PSI-BLAST.

	Experimento A			Experimento B								
COG AA	**Total EC**	**PSI+ SP_EC**	**PSI+ SP_EC+**	**PSI+EC/ PSI-EC**	**VP** PSI+ EC+	**FN** EC+	**VN** EC-	**FP** PSI+ EC-	**Precisão (%)**	**Revocação EC (%)**	***F-Score***	**MCC**
COG0013 Ala	3823	715	715	3630 / 0	3630	193	44.437	0	100	95,0	0,97	0,97
COG0017 Asp	4261	748	740	3841 / 4	3838	420	43.999	3	0,99	90,1	0,95	0,94
COG0017* Asn	2393	286	286	2312 / 0	2310	81	45.867	2	0,99	97,5	0,98	0,98
COG0018 Arg	3992	739	739	3829 / 0	3829	163	44.268	0	100	96,9	0,98	0,98
COG0060 Ile	3951	573	573	3736 / 0	3736	215	44.309	0	100	95,5	0,97	0,97
COG0124 His	4053	696	696	3727 / 0	3727	326	44.207	0	100	92,0	0,96	0,96
COG0143 Met	3262	496	496	2973 / 0	2973	289	44.998	0	100	91,1	0,95	0,95
COG0162 Tyr	4103	545	545	3964 / 0	3964	139	44.157	0	100	97,6	0,98	0,98
COG0172* Ser	3944	817	815	3740 / 0	3740	204	44.316	0	100	95,8	0,97	0,97
COG0180 Trp	2874	166	166	2781 / 0	2781	93	45.386	0	100	97,8	0,98	0,98
COG0441* Thr	4051	699	699	3889 / 0	3888	162	44.209	1	0,99	96,6	0,98	0,98
COG0495* Leu	3722	764	761	3609 / 0	3607	113	44.538	2	0,99	97,9	0,98	0,98
COG0525 Val	3831	343	335	3622 / 0	3621	209	44.429	1	0,99	94,5	0,97	0,97
TOTAL	48.260	7.487	7.466 (99,7%)	45.644 / 4	-	-	-	-	-	-	-	-

Experimento A: *Seeds* de aminoacil t-RNA sintetases foram comparadas com o banco de dados completo; Experimento B: *Seeds* de aminoacil t-RNA sintetases foram comparadas somente com o banco de dados contendo somente as 13 aminoacil t-RNA sintetases; SP: SwissProt; AA: aminoácido; Total EC: número total de proteínas da base estudada com o número EC associado da aminoacil-tRNA sintetase correspondente; SP_EC: número total de proteínas agrupadas que estão na base SwissProt com algum número EC associado; EC+: número EC esperado, idêntico ao presente nas *Seeds*. EC-: número EC diferente do esperado; PSI+: PSI validado; PSI-: não validado; PSI+EC: total de sequências PSI-validadas com algum número EC; VP: verdadeiro positivo; FN: falso negativo; VN: verdadeiro negativo; FP: falso positivo; Precisão: PSI+EC+/PSI+EC; Revocação: PSI+EC+/Total EC; *F-Score*: calculado por 2x (P x R) / (P + R); MCC: *Matthews correlation coefficient*, calculado por (TP x TN) – (FP x FN) / RAIZ((TP+FP) (TP+FN) (TN+FP) (TN+FN)); EC%: porcentagem de sequências no agrupamento associadas a algum número EC; *COGs sub-divididos pelo Seed Linkage.

6.1.1.4. Correlação entre família PANTHER, número EC e validação PSI-BLAST

Na seção anterior vimos a concordância existente entre a validação baseada em PSI-BLAST e anotações de número EC. Para avaliar esse processo de validação do SeedServer com uma metodologia diferente, foi escolhida a classificação de sequências nas melhores famílias e subfamílias PANTHER, grupos filogenéticos curados e supervisionados (ver métodos, seção 5.1.2).

Inicialmente, para verificar a correlação existente entre a classificação em grupos PANTHER com números EC, todas as aminoacil-tRNA sintetases possuidoras de uma anotação EC (Tabela 4, coluna 2) foram submetidas a essa classificação, sendo verificada se a designação nos níveis de família e subfamília estavam de acordo com a função enzimática EC correspondente (Tabela 5). Em alguns casos o teste foi impossibilitado pela ausência de uma descrição PANTHER específica (e.g.: família não nomeada). A análise gerou valores de *F-Score* e MCC para ambos os níveis, sendo todos superiores a 0,96 para o nível família.

Já para o nível de subfamília a precisão foi superior a 71,6 em cinco de oito testes realizados, mas alcançou baixos valores como 29,8 para treonil-tRNA sintetase. No entanto, para todos os casos (incluindo os níveis de família e subfamília) a presença de grupos carentes de descrição específica interfere na obtenção da medida real uma vez que esses foram considerados falsos positivos, como, por exemplo, é o caso de: i- treonil-tRNA sintetase: 69,9% das sequências classificadas em subfamílias PANTHER eram em grupos carentes de descrição e ii- metionil-tRNA sintetase: 63,8% das sequências classificadas pelo PANTHER eram em subfamílias carentes de descrição. Adicionalmente, a baixa atribuição do

classificador PANTHER em subfamílias contribuiu para baixos valores de revocação, como 2,1 para tirosil-tRNA sintetase. A combinação de baixos valores de precisão e revocação levou a baixos valores de *F-Score* e MCC no nível de subfamília para alguns testes, no entanto, sempre que a maioria das sequências estavam classificadas em grupos com descrição específica, a precisão foi alta independentemente da revocação.

Em uma próxima etapa, analisamos a correlação existente concomitantemente entre o total de proteínas PSI-validadas com número EC e classificação PANTHER. Nas colunas 8 e 9 da Tabela 5 fica evidente a correspondência da validação SeedServer com ambas as metodologias testadas, limitada somente pela baixa classificação PANTHER no nível de subfamília. Levando-se em consideração todas as proteínas PSI-validadas, independentemente de ter ou não uma atribuição EC, alguma família pode ser atribuída para 99,85% delas com descrição especificada e desse total 86,94% foi correspondente com o esperado para cada COG. Enquanto isso, alguma subfamília foi atribuída a 80,42% das sequências, sendo 53,96% com descrição especificada e desse total 91,04% de acordo com o esperado. Já para as não PSI-validadas, o índice de atribuição foi de aproximadamente 82% e 67% para família e subfamília respectivamente. Os casos onde as sequências não validadas possuem termos PANTHER corretos serão discutidos adiante e se justificam pelo pequeno tamanho das sequências e improbabilidade de atingirem o limiar de corte.

Para avaliar aquelas proteínas PSI-validadas que não apresentam uma correlação pré-estabelecida com algum número EC, analisamos a correlação entre essas sequências e grupos PANTHER. Nas duas últimas colunas da Tabela 5 verificamos alta correspondência no nível de família e menor para subfamília como já discutido.

Tabela 5: Correlação entre família PANTHER, número EC e validação PSI-BLAST.

COG AA	Família PANTHER / EC				Subfamília PANTHER / EC				PSI+ EC PANTHER		Proteínas PSI+ sem número EC	
	Precisão	Revocação	*F-Score*	MCC	Precisão	Revocação	*F-Score*	MCC	PSI+EC+ PTHR+	PSI+EC+ PTHRSF+	PTHR+	PTHRSF+
COG013 Ala	99,6	98,9	0,99	0,99	N.D.	N.D.	N.D.	N.D.	0,99	N.D.	100	N.D.
COG0017 Asp	99,7	92,0	0,96	0,96	93,8	85,7	0,89	0,92	0,96	92,6	100	41,3
COG0017* Asn	N.D.	N.D.	N.D,	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.
COG0018 Arg	99,9	98,0	0,99	0,99	N.D.	N.D.	N.D.	N.D.	0,99	N.D.	100	N.D.
COG0060 Ile	97,4	96,4	0,97	0,98	71,6	69,2	0,70	0,81	0,98	71,7	100	56,7
COG0124 His	99,8	97,2	0,98	0,98	N.D.	N.D.	N.D.	N.D.	0,98	N.D.	99,8	N.D.
COG0143 Met	97,7	93,6	0,96	0,96	35,1	33,2	0,34	0,53	0,96	35,6	100	69,0
COG0162 Tyr	99,7	99,0	0,99	0,99	93,5	2,1	0,04	0,13	0,99	94,6	98,1	93,2
COG0172* Ser	99,1	95,6	0,97	0,98	N.D.	N.D.	N.D.	N.D.	0,98	N.D.	99,9	N.D.
COG0180 Trp	99,9	99,0	0,99	0,99	99,9	99,0	0,99	0,99	0,99	98,1	99,8	99,7
COG0441* Thr	N.D.	N.D.	N.D.	N.D.	29,8	29,2	0,29	0,51	N.D.	29,8	N.D.	66,2
COG0495* Leu	99,9	99,1	0,99	0,99	99,8	98,9	99,4	0,99	0,99	99,4	100	96,1
COG0525 Val	99,7	96,8	0,98	0,98	54,9	53,3	0,54	0,71	0,98	54,4	100	71,1

AA: aminoácidos; PSI+: sequências PSI-validadas; EC+: sequências com o número EC esperado; PTHR: família PANTHER; PTHRSF: subfamília PANTHER; PTHR+: sequências com a família PANTHER esperada; PTHRSF+: sequências com a subfamília PANTHER esperada; PSI+ EC PANTHER: total de sequências PSI-validadas com alguma atribuição EC e alguma atribuição PANTHER; Precisão, Revocação, *F-Score* e MCC: como descritos na Tabela 4; N.D.: Dados não disponíveis por ausência de descrição PANTHER específica; *COGs divididos pelo Seed Linkage.

Finalmente, a atribuição PANTHER completa para proteínas PSI-validadas se encontra na Tabela 6, as descrições dos termos se encontram na legenda, quando há coincidência com a descrição do COG os identificadores foram representados em negrito e itálico e quando desprovidos de descrição marcados em vermelho. Denotamos a seguir casos onde houve conflito de anotação.

Aspartil e Asparaginil-tRNA sintetases foram classificados na mesma família PTHR22594 (Aspartyl/Lysyl-tRNA Synthetase) e, curiosamente, não há nessa versão da base de dados PANTHER uma família Asparaginil-tRNA sintetase além do fato de suas estruturas primárias [43] e terciárias [44] serem altamente similares. Das quatro sequências não PSI-validadas, mas com número EC correto (Tabela 5), três (Q6MTH6, F8K681 e C4QJ96) possuem tamanho inferior a 100 resíduos de aminoácidos e somente uma Q5FNR4 com 311 resíduos, porém nenhuma família PANTHER lhe foi atribuída.

Alguns casos de troca nas subfamílias da família PTHR11946, que agrupa Isoleucil, Leucil e Valil-tRNA sintetases, estão presentes como esperado pelo que já foi descrito no item anterior. A troca mais expressiva foi a presença no COG0525 de Valil-tRNA sintetase de 420 sequências da subfamília PTHR11946-SF16 que é relacionada a Ile-tRNA sintetase. No entanto, para essas 420 sequências foram analisados os resultados secundários dentro dos critérios de *E-value* estabelecidos pelo programa *PANTHER HMM Scoring tool* (ver métodos, seção 5.1.2) e 358 delas apresentam correlação com a subfamília esperada PTHR11946-SF5, PTHR11946-SF5 (Valil-tRNA sintetase).

Para Histidil-tRNA sintetase, a já mencionada D8S8H4 contendo 1.665 resíduos de aminoácidos e PSI-validada foi classificada como PTHR18934, uma RNA helicase. No entanto essa proteína é multifuncional e também apresenta atividade de

Histidil-tRNA sintetase, em suporte ao determinado pelo SeedServer. Já as outras duas sequências (D8SPU1 e D8SXN2 com respectivamente 1.142 e 935 resíduos) não foram validadas e possuem unicamente função de RNA helicase.

Para Tirosil-tRNA sintetase, 16 sequências foram atribuídas a PTHR10055-SF0 (Triptofanil-tRNA sintetase). As cinco sequências não PSI-validadas atribuídas corretamente a PTHR11766 de Tirosil-tRNA sintetase possuem apenas 78 resíduos de aminoácidos, se comparados com uma média de 340 das *Seeds*.

6.1.1.5. Enriquecimento taxonômico

O enriquecimento taxonômico ocorrido devido ao recrutamento SeedServer foi avaliado para todas as sequências das treze aaRSs, que foram consideradas PSI-validadas, e com o número EC e família PANTHER corretamente atribuídos. Proteínas de novos táxons foram encontradas, representando até mesmo novos filos (como os bacterianos *Lentisphaerae, Poribacteria*, *Thermodesulfobacteria* e eucarióticos *Phaeophyceae* e *Platyhelminthes*) e centenas de novas espécies comparando-se com o caso de se somente grupos KO tivessem sido utilizados (Tabela 7). O que torna o resultado expressivo é o fato de que se espera fácil identificação de aaRSs presentes em outros clados e inclusão das mesmas nos grupos KO.

Tabela 6: Correlação entre famílias e subfamílias PANTHER e validação PSI-BLAST.

COG AA	PSI+/PSI-	PTHR/PSI+	PTHRSF/PSI+	PTHR/PSI-	PTHRSF/PSI-
COG0013Ala	4160 / 0	***PTHR11777*** (4160)	PTHR11777-SF6 (4099)	-	-
COG0017Asp	4715 / 8	***PTHR22594*** (4715)	***PTHR22594-SF5*** (3913) PTHR22594-SF10 (706) PTHR22594-SF16 (47)	PTHR22594 (4)	PTHR22594-SF5 (4)
COG0017*Asn	2864 / 0	PTHR22594 (2860) PTHR10188 (1)	PTHR22594-SF6 (1716) PTHR22594-SF16 (1093) PTHR22594-SF18 (51) PTHR10188-SF7 (1)	-	-
COG0018Arg	4317 / 0	***PTHR11956*** (4313)	PTHR11956-SF0 (397) PTHR11956-SF1(2770)	-	-
COG0060 Ile	4375 / 0	***PTHR11946*** (4373)	***PTHR11946-SF9*** (3041) PTHR11946-SF11 (1317) PTHR11946-SF41 (15)	-	-
COG0124 His	4339 / 2	***PTHR11476*** (4331) PTHR18934 (1)	PTHR18934-SF68 (1)	***PTHR18934*** (2)	PTHR18934-SF68 (2)
COG0143 Met	4552 / 0	***PTHR11946*** (4551)	***PTHR11946-SF1*** (2140) PTHR11946-SF13 (631) PTHR11946-SF47 (1738) PTHR11946-SF48 (39)	-	-
COG0162 Tyr	4715 / 6	***PTHR11946*** (474) ***PTHR11766*** (4223) PTHR10055 (16)	PTHR11946-SF1 (1) ***PTHR11946-SF8*** (308) PTHR11946-SF45 (1) PTHR10055-SF0 (16)	PTHR11946 (1) PTHR11766 (5)	-
COG0172* Ser	4555 / 0	***PTHR11778*** (4490) PTHR11451 (1) PTHR31084 (1)	PTHR11778-SF0 (355) PTHR11778-SF1 (3315) PTHR31084-SF0 (1)	-	-
COG0180 Trp	4865 / 6	***PTHR10055*** (4862)	***PTHR10055-SF0*** (4204)	***PTHR12363*** (5)	PTHR12363-SF13 (5)

		PTHR11700 (2) PTHR24115 (1)	***PTHR10055-SF1*** (425) ***PTHR10055-SF2*** (51) PTHR11700-SF1 (2) PTHR24115-SF227 (1)	***PTHR11700*** (1)	PTHR11700-SF1 (1)
COG0441* Thr	4484 / 0	PTHR11451 (4483)	***PTHR11451-SF5*** (1553) PTHR11451-SF10 (2913) PTHR11451-SF11 (16)	-	-
COG0495* Leu	4362 / 0	***PTHR11946*** (4362)	PTHR11946-SF5 (1) ***PTHR11946-SF6*** (149) ***PTHR11946-SF7*** (3907) PTHR11946-SF9 (1) ***PTHR11946-SF10*** (256) PTHR11946-SF11 (1) PTHR11946-SF20 (9) PTHR11946-SF21 (20) PTHR11946-SF24 (2) PTHR11946-SF49 (16)	-	-
COG0525 Val	4199 / 0	***PTHR11946*** (4199)	***PTHR11946-SF5*** (2378) PTHR11946-SF9 (18) PTHR11946-SF16 (420) PTHR11946-SF49 (1383)	-	-
TOTAL	56.502 / 22	56419 (99,85%)	45438 (80,42%)	18 (81,82%)	12 (66,67%)

AA: aminoácido; **PSI+**: Sequências PSI-validadas; **PSI-:** Sequências não PSI-validadas; **PTHR**: identificador da família PANTHER, **PTHRSF**: identificador da subfamília PANTHER; Total de sequências apresentado entre parênteses; *COGs divididos pelo Seed Linkage; Código de três letras dos aminoácidos foi utilizado; Termos PANTHER em negrito e itálico estão de acordo com o esperado para a descrição do COG; Termos PANTHER em vermelho não possuem descrição especificada. **PTHR11777**: Ala-tRNA sintetase - SF6:subfamília não nomeada; **PTHR22594**:Asp/Lys-tRNA sintetase - SF5:Asp-tRNA sintetase, SF6:Asp-tRNA sintetase, SF10:subfamília não nomeada, SF16:subfamília não nomeada; SF18: :subfamília não nomeada, **PTHR10188**:L-Asparaginase - SF7:L-Asparaginase; **PTHR11956**:Arg-tRNA sintetase - SF0:subfamília não nomeada, SF1:subfamília não nomeada; **PTHR11946**:Ile, Leu, Tyr, Val e Met-tRNA sintetases – SF1:Met-tRNA sintetase, SF5:Val-tRNA sintetase, SF6:relacionado a Leu-tRNA sintetase, SF7:Leu-tRNA sintetase, SF8:Tyr-tRNA sintetase, SF9:Ile-tRNA sintetase, SF10:Leu-tRNA sintetase, SF11:subfamília não nomeada, SF13:Proteína OS03G0335600, SF16:relacionada a Ile-tRNA sintetase, SF20:subfamília não nomeada, SF21:tRNA sintetase de classe I putativa, SF24:subfamília não nomeada, SF41:subfamília não nomeada, SF45:subfamília não nomeada, SF47:subfamília não nomeada, SF48:subfamília não nomeada, SF49:subfamília não nomeada; **PTHR11476**:His-tRNA sintetase; **PTHR18934:** RNA helicase ATP-Dependente - SF68 subfamília não nomeada; **PTHR11766**:Tyr-tRNA sintetase; **PTHR10055**:Trp-tRNA sintetase - SF0 Trp-tRNA sintetase - SF1 Trp-tRNA sintetase citoplasmática, SF2 Trp-tRNA sintetase mitocondrial; **PTHR12363**:Transportina 3 e importina 13 - SF13 subfamília não nomeada; **PTHR11778**:Ser-tRNA sintetase - SF0:subfamília não nomeada, SF1:subfamília não nomeada; **PTHR11451**: relacionada a tRNA sintetase; **PTHR31084**:família não nomeada – SF0: subfamília não nomeada; **PTHR11700**: Proteína Ribossomal 30S membro da família S10 - SF1 Não caracterizada; **PTHR24115**:Família não nomeada - SF227 Proteína OS02G0742800; **PTHR11451**:relacionado a tRNA sintetase- SF5 Thr-tRNA sintetase, SF10: subfamília não nomeada, SF11: subfamília não nomeada.

Tabela 7: Enriquecimento taxonômico como o recrutamento SeedServer das treze aminoacil-tRNA sintetases com relação ao que é encontrado no grupo KO apropriado.

COG AA	Filo	Classe	Ordem	Família	Gênero	Espécie
COG0013 Ala	3	4	7	23	179	807
COG0017 Asp	4	5	8	23	177	836
COG0017* Asn	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.
COG0018 Arg	3	5	8	25	183	815
COG0060 Ile	4	4	8	24	179	803
COG0124 His	2	5	9	26	173	783
COG0143 Met	2	2	5	18	141	606
COG0162 Tyr	4	5	7	24	180	832
COG0172* Ser	3	6	9	27	172	738
COG0180 Trp	3	5	8	23	142	514
COG0441* Thr	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.
COG0495* Leu	4	5	9	26	169	806
COG0525 Val	2	4	8	21	166	786

Distribuição de novos grupos em diferentes níveis taxonômicos das proteínas recrutadas pelo SeedServer, considerando PSI-validadas, número EC e família PANTHER corretamente atribuídas e ausência no grupo KO correspondente. N.D.: Dados não disponíveis por ausência de descrição PANTHER específica;*COGs divididos pelo Seed Linkage.

6.1.2. Metilaspartato mutase e subgrupos KO

Pelo menos uma via bioquímica para a fermentação dos 20 aminoácidos básicos é conhecida. A glutamato mutase (EC: 5.4.99.1) ou metilaspartato mutase (Mut) é uma enzima procariótica dependente de vitamina B12, que cataliza a conversão reversível de L-glutamato para L-treo-3-metilaspartato, o primeiro passo na fermentação do L-glutamato até piruvato. Mut possui duas subunidades distintas descritas: MutS – domínio de ligação a vitamina B12; MutE – a subunidade catalítica, além de MutL – cujo gene está localizado entre as duas outras subunidades e possui função desconhecida.

Considerando que essas três subunidades estão agrupadas em um mesmo grupo KO (K01846), a capacidade do SeedServer de separá-las foi avaliada, demonstrando facilidade em distinguir as subunidades e reagrupá-las em subgrupos distintos. A Figura 9 mostra uma árvore filogenética construída com todas as 51 sequências bacterianas presentes nesse KO e que foram utilizadas como *Seeds* no SeedServer. Os três subgrupos observados na árvore são os mesmos formados no agrupamento SeedServer pelo programa Seed Linkage. Por estarem inseridas em uma mesma entidade KO, as mesmas sequências do KO são recrutadas para cada um dos subgrupos formados durante o procedimento SeedServer, no entanto o processo de validação PSI-BLAST eliminou, para cada subgrupo, sequências das outras duas subunidades com 100% de acerto.

MutS formou um grupo com 167 proteínas, sendo 5,4%, 23,4% e 71,2%, respectivamente, das categorias PSI-Validadas: S, K e U (Figura 8).

MutE formou um grupo com 181 proteínas, sendo 21,6%, 8,8% e 69,6%, respectivamente, das categorias S, K e U.

Já MutL formou um grupo com 448 proteínas, sendo 35,9%, 9,6% e 54,5%, respectivamente, das categorias S, K e U, mostrando mais uma vez a importância do enriquecimento de sequências por parte do Seed Linkage e UEKO.

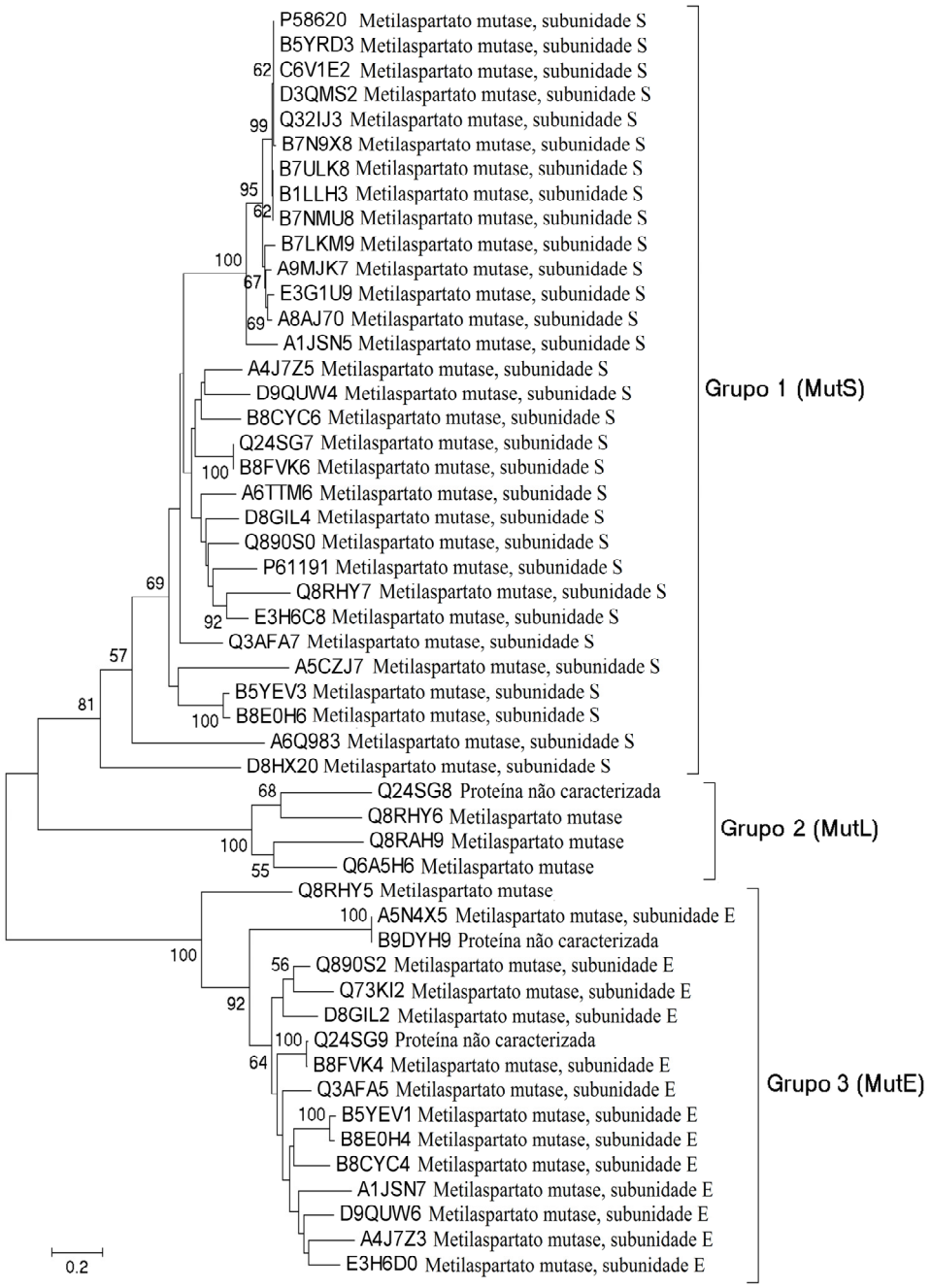


Figura 9: Árvore filogenética das 51 metilaspartato mutases presentes no K01846. MutS, MutL e MutS são suas três subunidades.

6.1.3. Vitamina B7 e anotação de sequências metagenômicas

Nessa seção, selecionamos uma enzima bacteriana envolvida no metabolismo de aminoácidos de forma aleatória para simular um processo de anotação de proteínas de função desconhecida a partir dos agrupamentos formados pelo SeedServer. Biotina ou vitamina B7 é uma coenzima para enzimas carboxilases, envolvida na síntese de ácidos graxos, gliconeogênese e dos aminoácidos isoleucina, leucina e valina. A proteína bifuncional de síntese de biotina contém dois domínios funcionais. Três sequências (Q64VX4, Q5LEY1 e E1WTS4) de uma bactéria do trato intestinal, *Bacteroides fragilis,* foram utilizadas como *Seeds* na criação de um grupo de homólogos.

Um total de 5.838 proteínas recrutadas, dentre procarióticas e eucarióticas, provenientes das categorias PSI-validadas S, K, U, P e Q foram então utilizadas em uma busca BLAST (*E-value* 10^{-10}) usando como banco de dados 18.445.837 sequências públicas de projetos metagenômicos, com o intuito de associar informação biológica a essas sequências.

Como algumas sequências metagenômicas, após a montagem, podem conter milhares de pares de base, foi feita a seleção do melhor resultado BLAST por região, ou seja, sem sobreposição. Em outras palavras, os melhores alinhamentos de sequências UniProtKB puderam ser mapeadas em diferentes regiões de uma única sequência metagenômica.

Após a filtragem, foi feita a verificação de quais categorias SeedServer foram responsáveis pelos melhores alinhamentos. De um total de 1.676 alinhamentos, 29,1% e 70,9% foram respectivamente do UEKO (sua porção exclusiva) e KO. Se somente as sequências *Seeds* fossem utilizadas no BLAST, somente 336 alinhamentos seriam obtidos. Esse resultado mostra mais uma vez a importância da

utilização do UEKO como fonte de uma melhor anotação para sequências de função desconhecida.

6.1.4. Via regulatória do desenvolvimento pré-embrionário e inferência de LCA

Neste próximo exemplo desejamos ilustrar o uso da ferramenta SeedServer em estudos mais complexos envolvendo vias metabólicas completas e não somente reações isoladas. Em um trabalho [45] desenvolvido por nosso grupo ferramentas de mineração de texto foram utilizadas para encontrar proteínas envolvidas na regulação da via de desenvolvimento pré-embrionário humano descrevendo uma nova via bioquímica, no mesmo estilo encontrado no KEGG *Pathway*. As sequências protéicas foram então utilizadas como *Seeds* no SeedServer permitindo a inferência do ancestral comum mais recente (LCA na sigla em inglês) para cada grupo de homólogos, revelando que a via consiste de conjuntos de interações conservadas no filo dos cordados, mas com importante adição de elementos mais recentes.

Como ilustrado na Tabela 7 acima, o SeedServer é responsável por recrutar proteínas homólogas de táxons novos e esse enriquecimento pode levar ao reposicionamento do LCA, influenciando assim na demarcação da origem do gene. A Figura 10 demonstra essa remarcação do LCA ocorrida com o uso do SeedServer para alguns (pontos que não estão sobre a diagonal tracejada) dos 115 grupos de homólogos criados para via em questão, quando comparado ao uso exclusivo da base de dados KO. Como vários grupos podem ocorrer com as mesmas coordenadas, o tamanho das bolhas representa a freqüência de grupos naquela coordenada. Quando o grupo formado pelo SeedServer agregou sequências de mais

clados, o LCA foi mapeado abaixo da diagonal, em um clado mais remoto. Esse estudo de caso exemplifica a necessidade de um usuário em rapidamente ampliar ou criar um grupo de homólogos com todas as proteínas disponíveis no momento.

Em alguns casos, quando as proteínas selecionadas não estavam presentes no KO, a determinação do LCA foi integralmente determinada pelo SeedServer, através do Seed Linkage e do método UE-*Seed* (pontos sobre o eixo Y da Figura 10). Por exemplo, as sequências SwissProt P61158 (similar a actina 3) e O75626 (proteína contendo o domínio dedos de zinco PR) não estão presentes em nenhum grupo KO até a presente data, mas tiveram seus homólogos determinados.

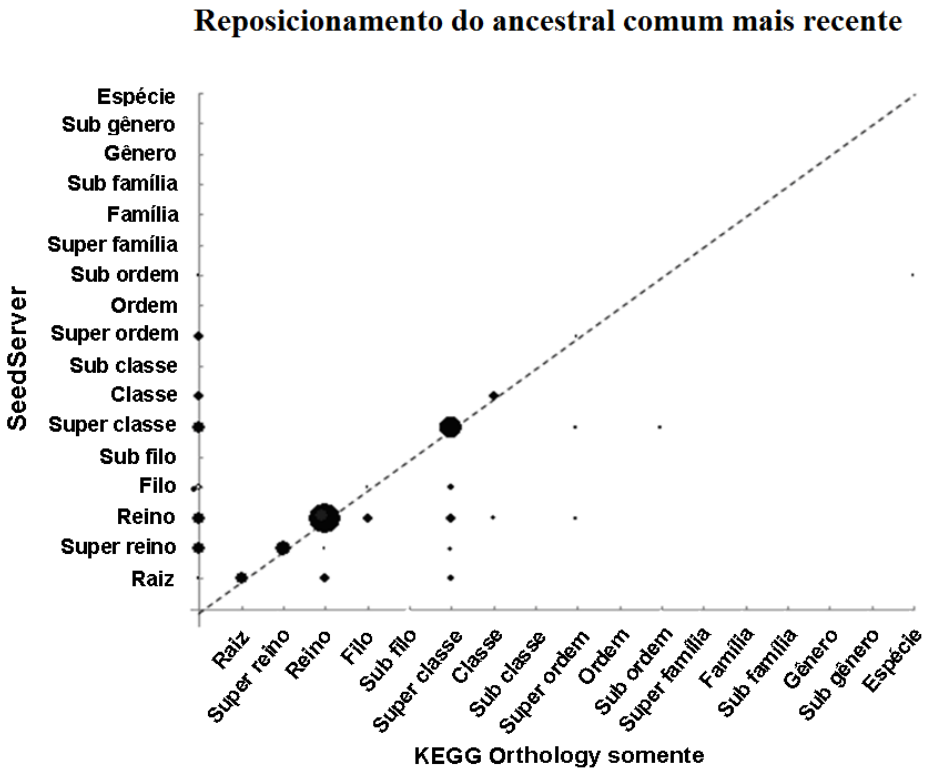


Figura 10: Reposicionamento do ancestral comum mais recente (LCA) com uso do SeedServer. Dados de todas as proteínas utilizadas como *Seeds* no estudo da via de regulação do desenvolvimento pré-embrionário humano. Eixo X considera somente sequências presentes em grupos KO e eixo Y todas sequências ao final do processamento SeedServer. Tamanho das bolhas é proporcional ao número de grupos em cada coordenada. Valor mínimo = 1 e máximo = 32.

6.1.5. LCA e estrutura secundária

A disponibilidade de milhares de dados contendo a porcentagem de similaridade das estruturas secundárias (SOV) entre *Seeds* e recrutadas, aliado à condição de se estabelecer uma distância taxonômica de cada par, através do LCA, levantou as seguintes questões: (i) Quão similares são as estruturas secundárias em diversos níveis taxonômicos dentre sequências dos três diferentes super-reinos da vida?; (ii) Seria possível estabelecer classes de valores SOV que distinguem esses níveis taxonômicos?

Para abordar essas questões, juntamente com os dados dos agrupamentos já apresentados, foi feito o agrupamento de homólogos de sequências de mioglobinas, com estruturas sabidamente conservadas a título de comparação e controle, utilizando como *Seeds* uma sequência humana (P02144) e uma de camundongo (P04247). Além destes, somam-se os dados que representam o grupo das plantas, utilizando genes de mais uma via completa mapeada por nosso grupo (Fernanda Stussi e colaboradores, não publicado), envolvendo proteínas de resistência à seca (*Seeds* de *Arabidopsis thaliana*).

A Figura 11 nos mostra dados de SOV para procariotos, archaeas e eucariotos distribuídos em até quinze níveis taxonômicos. De maneira geral é possível observar pela inclinação positiva de todas as retas a já esperada correlação de quanto menor a distância taxonômica maior será o valor SOV, independentemente do reino estudado.

Analisando-se em conjunto os dados de Aspartato mutase (procariótico), aminoacil-tRNA sintetases (archaeas) e proteínas do desenvolvimento pré-embrionário humano (eucariotos metazoários) observamos uma alta correlação entre as retas, um indicativo da conservação da estrutura secundária em torno de 70% de

proteínas homólogas entre os três super-reinos, com elevação similar até cerca de 90% próximo ao nível de espécie.

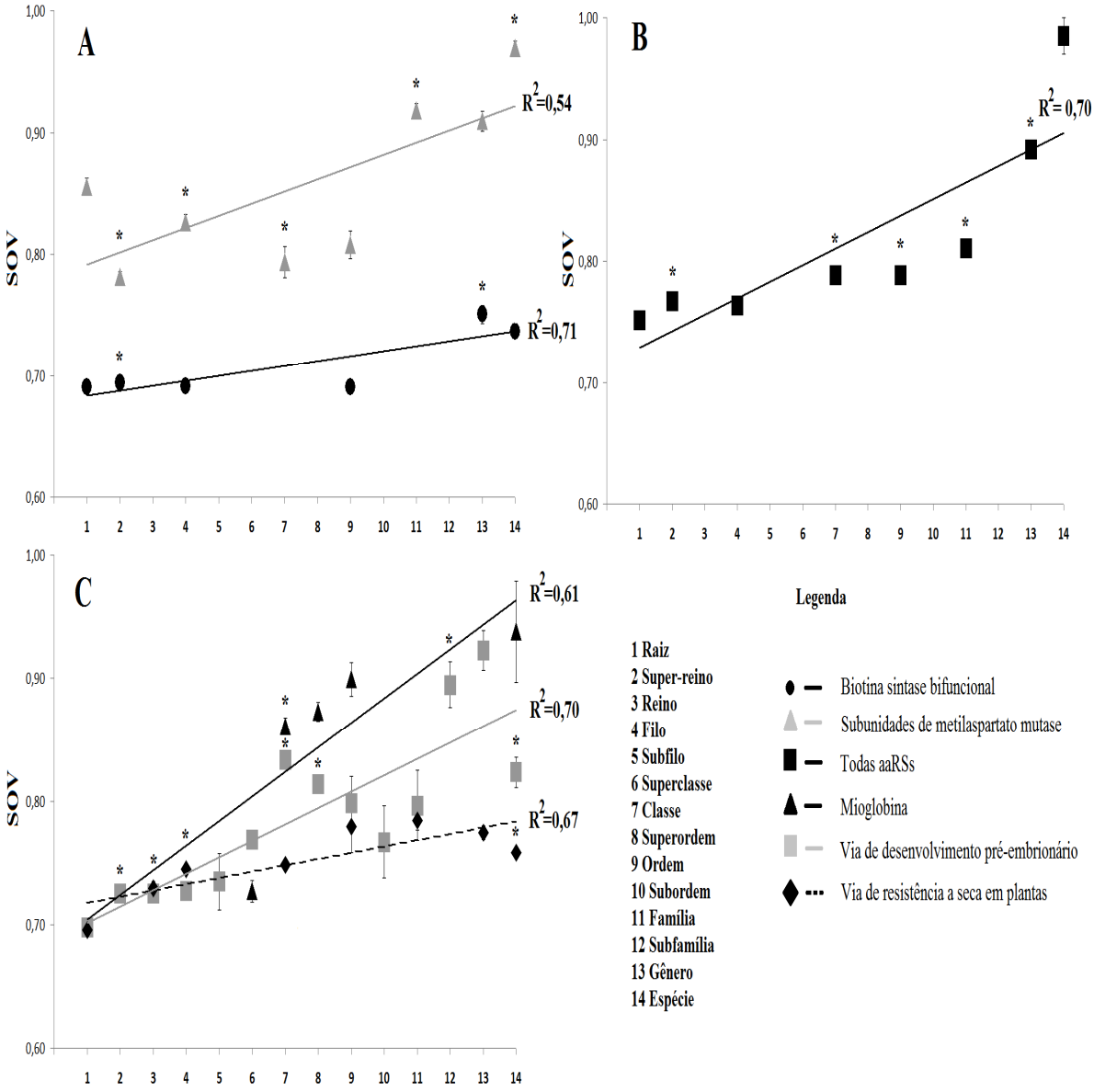


Figura 11: Correlação da sobreposição de estruturas secundárias dada pelo valor SOV (eixo Y) com distância taxonômica (eixo X) entre *Seeds* e todas as sequências PSI-validadas. A: grupos com *Seeds* procarióticas; B: *Seeds* de archaea; C: *Seeds* de eucariotos. *dados de um nível taxonômico que apresentam diferença estatística significativa em relação ao nível anterior em Teste-T não pareado, assumindo variâncias desconhecidas e *p-value* < 0,05. Barra de erros representa o desvio padrão da média.

Os valores de SOV e a inclinação obtida para Biotina sintase mostraram-se consideravelmente inferiores quando comparados a Aspartato mutase de outros procariotos bem como de archaea e eucariotos. No entanto, como as sequências *Seeds* utilizadas são bi-funcionais e estão associadas a dois grupos KO distintos, o processo de enriquecimento KO/UEKO recrutou sequências mono-funcionais de menor tamanho, fato que influencia na métrica de obtenção de SOV reduzindo os valores obtidos (Figura 11-A). Especialmente nesse experimento, valores de SOV baixos alertam para a mono e multifuncionalidade em grupos KO, sendo uma perspectiva futura a subdivisão de grupos KO análogos em distintos grupos mono e multifuncionais, para que cada subgrupo contenha proteínas similares ao longo de toda extensão.

A inclinação da reta obtida para mioglobina foi a maior, juntamente com os valores individuais de SOV refletindo a conservação dessas proteínas. Contrariamente, as proteínas de planta apresentaram menor inclinação e os menores valores SOV, resultado compatível com a presença de diversos parálogos nos genomas multiplóides e a provável falta de depósito dos homólogos mais próximos, o que incorre em maior variação estrutural (Figura 11-C).

É importante ressaltar que o eixo X dos gráficos contém unidades discretas e não contínuas como seria esperado em uma representação temporal. Nesse sentido, a separação entre os pontos não reflete a distância evolutiva que levou a separação dos táxons. Ainda, a elevada diferença entre alguns valores consecutivos, principalmente próximos ao nível de espécie de procariotos e archaeas, sugere uma divisão taxonômica incompleta, onde mais níveis deveriam ser criados para ajustar melhor o que seria a real distância evolutiva e, que seria representada com os pontos corretamente afastados entre si e posicionados sob a reta. Estas observações

sugerem a necessidade de estudos posteriores, todavia, as distâncias entre os clados em vários super-reinos são compatíveis.

Uma melhor visão da distribuição de valores SOV em clados diferentes está mostrada na Figura 12 para as proteínas envolvidas no desenvolvimento pré-embrionário humano. Verifica-se que a maioria das proteínas que somente têm em comum o filo com as humanas, apresentam valores de SOV mais baixos, enquanto que as mais próximas têm valores maiores, no entanto, observa-se certa sobreposição de valores entre os níveis taxonômicos, indicando a impossibilidade de estabelecer classes de valores SOV para diferenciar níveis taxonômicos. No SeedServer, o parâmetro SOV não tem caráter de filtro ou validação, embora isso possa ser facilmente aplicado à tabela de resultados.

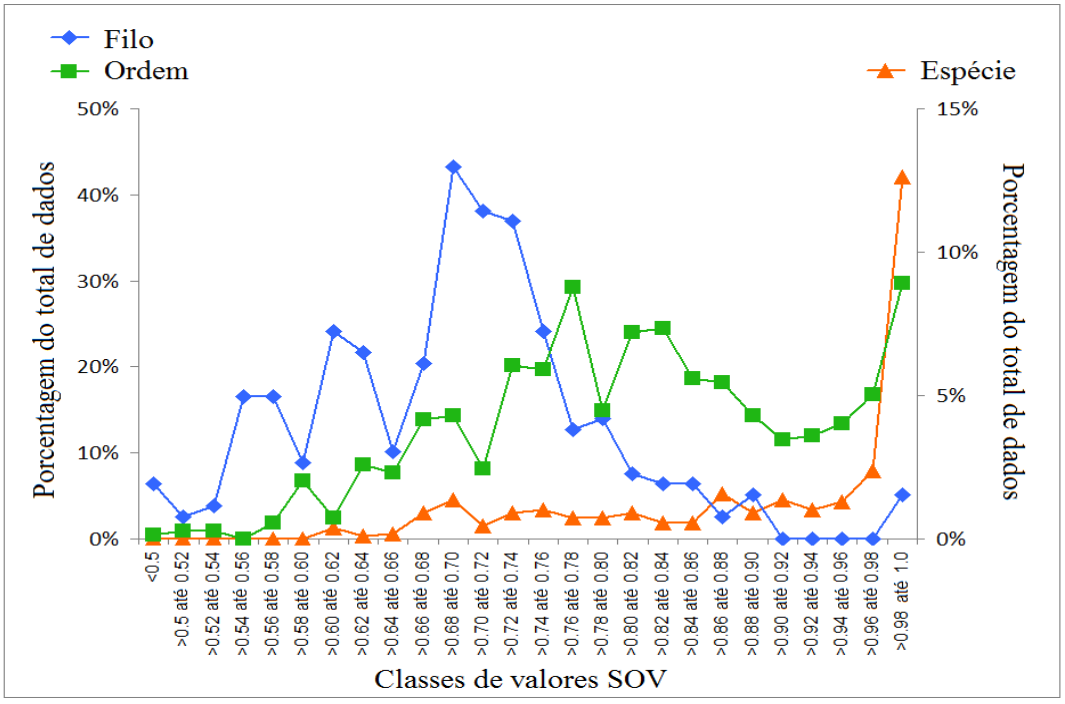


Figura 12: Distribuição em classes distintas de valores SOV obtidos em três níveis taxonômicos em razão do total de dados. Filo e ordem referentes ao eixo Y esquerdo e espécie ao eixo Y direito. Dados referentes aos grupos de homólogos formados para proteínas da via de desenvolvimento pré-embrionário humano.

6.1.6. Contribuição de genomas não completos e componentes SeedServer

A porcentagem de proteínas advindas de genomas com diferentes níveis de sequenciamento foi aferida para quatro estudos apresentados nesse trabalho. A Figura 13 revela que a participação de genomas ainda em processo de montagem, ou mesmo ainda sem um projeto de sequenciamento estabelecido, prevalece com mais de 50% do total de proteínas recrutadas nos grupos de homólogos. Adicionalmente, a Figura 14 resume o papel exclusivo de cada componente SeedServer (SeedLinkage, KO e UEKO) onde é possível constatar a predominância da porção UEKO em todos os casos. Evidentemente cada situação dependerá do tipo de estudo e da qualidade da anotação em grupos KO, mas como dispomos de casos diferentes, podemos apreciar as variações.

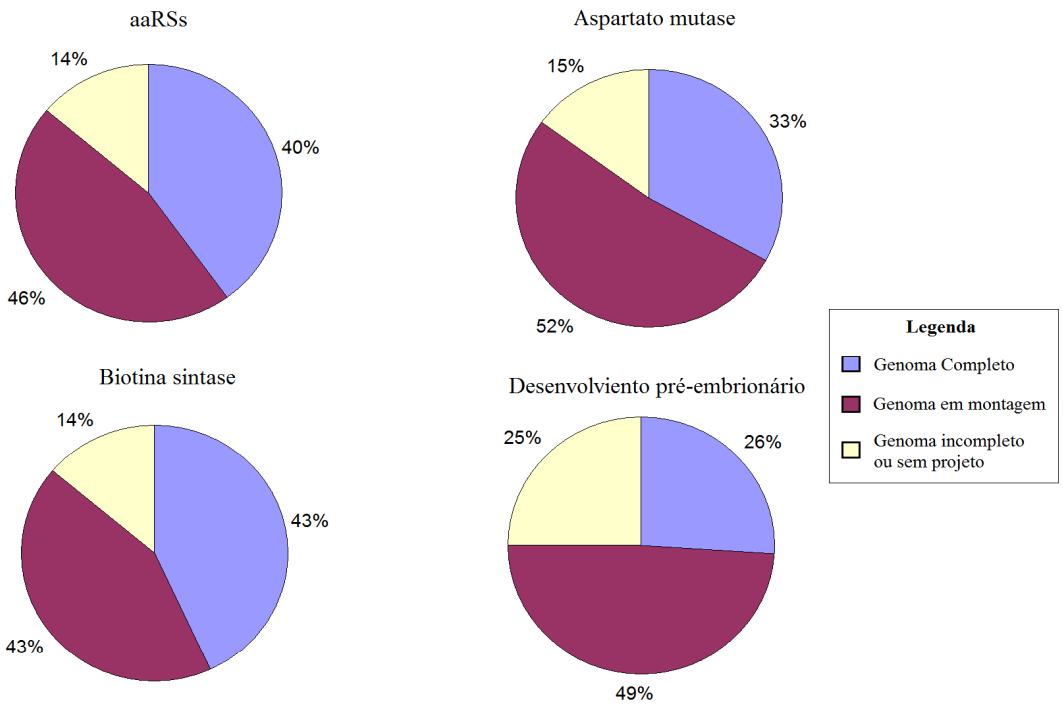


Figura 13: Participação de proteínas provenientes de genomas com diferentes níveis de sequenciamento. aaRS: todas aminoacil-tRNA sintetases estudadas.

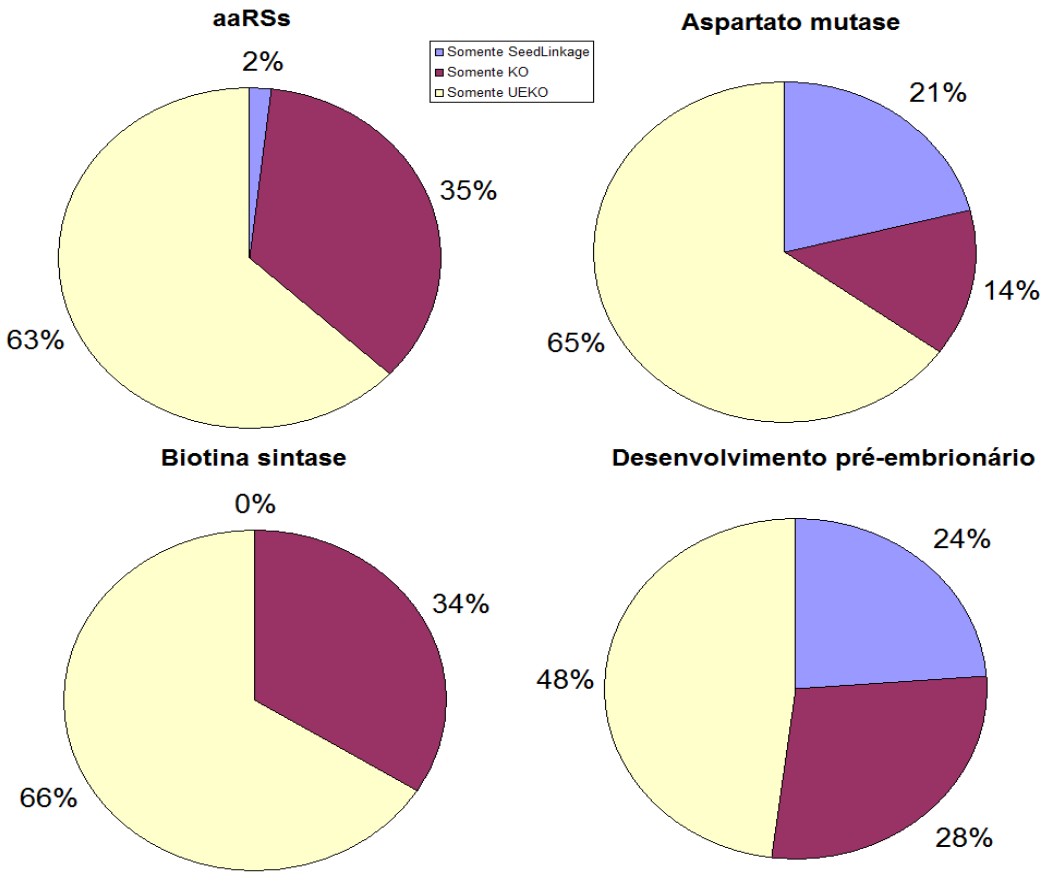


Figura 14: Participação de proteínas provenientes das diferentes porções do recrutamento SeedServer. aaRS: todas aminoacil-tRNA sintetases estudadas.

6.1.7. SeedServer em interface *web*

Para facilitar e incentivar a utilização do SeedServer foi criada uma interface *web* [http://pinguim.fmrp.usp.br/cenabid/form_SS.html] de fácil manuseio onde é necessário apenas o cadastramento do usuário sem a necessidade de senha e escolha de alguns parâmetros, sendo que assim que processados os resultados podem ser visualizados a qualquer momento.

Cada processo criado gera um identificador numérico único e uma página de consulta é disponibilizada para cada usuário informando a lista completa de *Seeds* respectivamente utilizadas, de forma a manter um histórico dos estudos já

realizados. Para auxiliar novos usuários foi criado um guia passo a passo completo, perfazendo todas as etapas do recrutamento SeedServer e ajuda na interpretação dos dados obtidos (Figura 15).

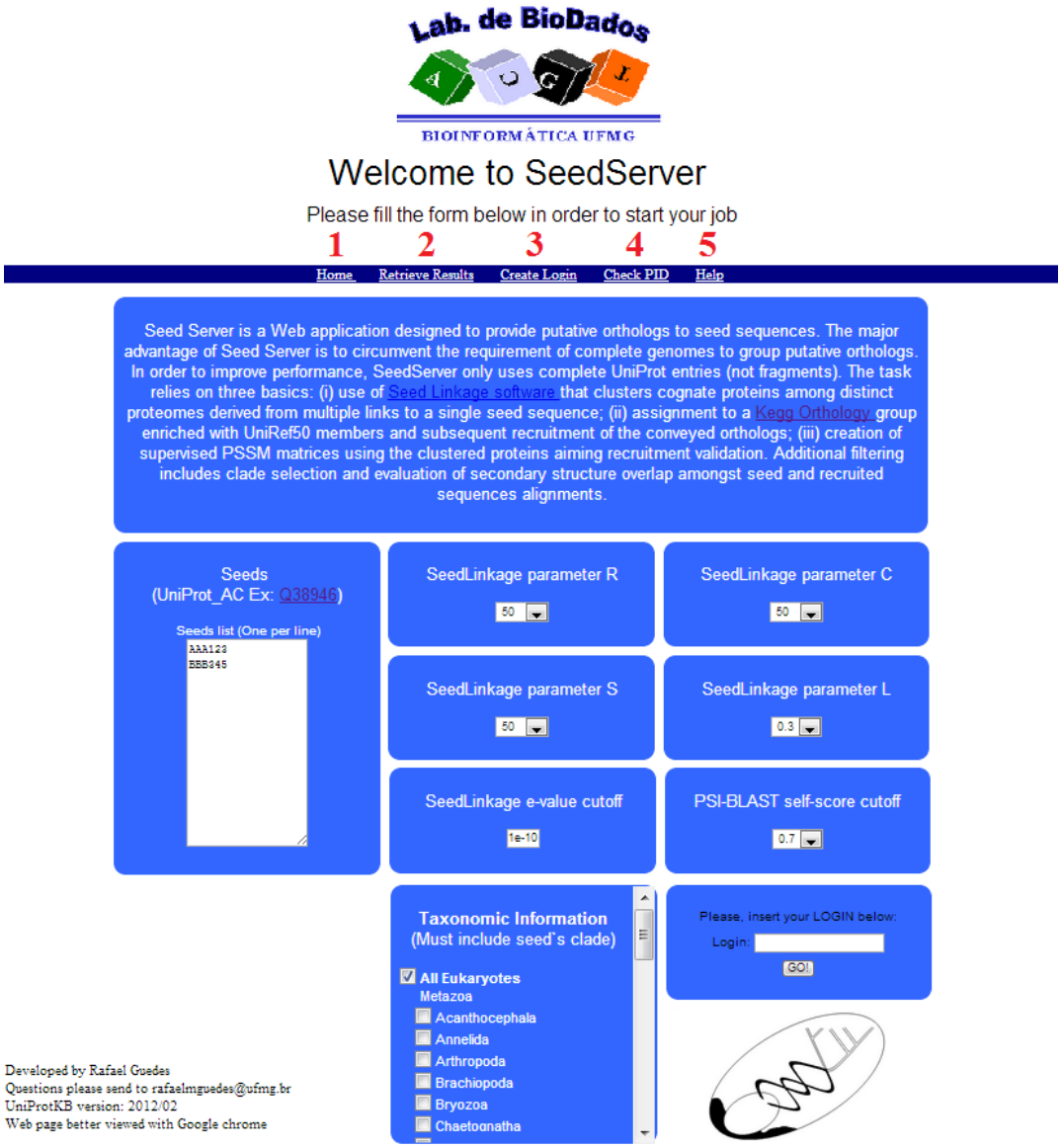
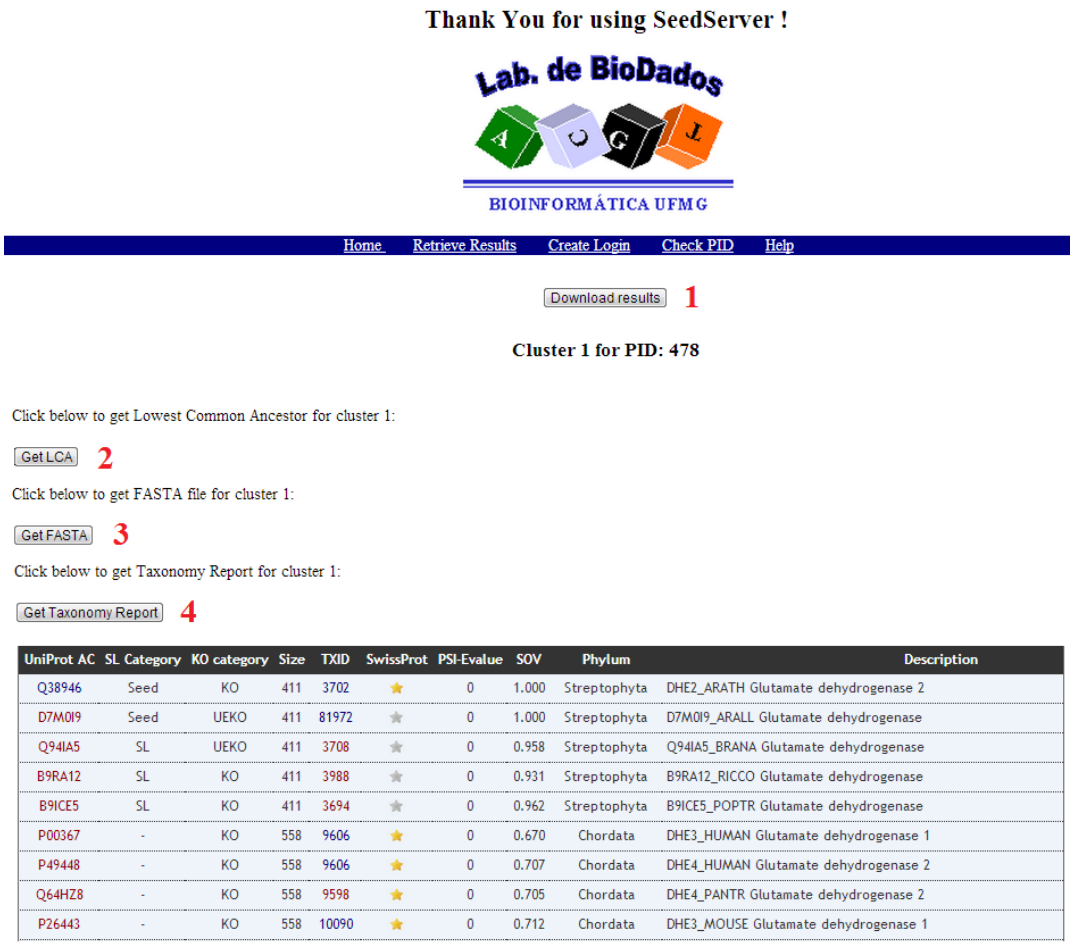


Figura 15: Página *web* principal desenvolvida para disponibilização do SeedServer. Atalhos representados por números em vermelho na Figura - 1: Página principal para escolha dos parâmetros, como *Seed*(s), parâmetros Seed Linkage, grupo taxonômico e validação PSI-BLAST; 2: Página para obtenção e visualização dos resultados; 3: Cadastramento de novos usuários; 4: Obtenção do histórico de projetos por usuário; 5: Guia de ajuda.

De posse de um determinado identificador numérico que identifica o processo o usuário efetua a consulta dos resultados SeedServer visualizados em forma de tabela (Figura 16).

Nesse momento é possível consultar dados referentes ao recrutamento para as sequências PSI-validadas como a categoria que a proteína pertence (Seed Linkage, KO e UEKO), número de resíduos de aminoácidos, identificador numérico taxonômico interligado à base *Taxonomy* do NCBI, informação de se a proteína é SwissProt ou TrEMBL, valores de *E-value* usados na validação PSI-BLAST, maior SOV obtido com a(s) *Seed*(s) e finalmente a descrição da sequência FASTA disponível pelo UniProtKB.

Esta página também oferece serviços adicionais, como disponibilização para descarregamento dos dados referentes ao agrupamento, obtenção detalhada do LCA, sequências em formato FASTA e, por fim, um relatório taxonômico completo. É importante antecipar que a determinação detalhada do LCA é realizada por uma requisição remota feita a outro servidor, que contém toda a tabela taxonômica em memória e, portanto pode retornar a informação precisa do clado mais recente que reúne todos os identificadores taxonômicos dos homólogos agrupados. No relatório taxonômico é utilizada uma tabela com somente os clados de Lineu, portanto algumas informações de LCA são referentes ao clado de Lineu mais próximo.



UniProt AC	SL Category	KO category	Size	TXID	SwissProt	PSI-Evalue	SOV	Phylum	Description
Q38946	Seed	KO	411	3702	★	0	1.000	Streptophyta	DHE2_ARATH Glutamate dehydrogenase 2
D7M0I9	Seed	UEKO	411	81972	☆	0	1.000	Streptophyta	D7M0I9_ARALL Glutamate dehydrogenase
Q94IA5	SL	UEKO	411	3708	☆	0	0.958	Streptophyta	Q94IA5_BRANA Glutamate dehydrogenase
B9RA12	SL	KO	411	3988	☆	0	0.931	Streptophyta	B9RA12_RICCO Glutamate dehydrogenase
B9ICE5	SL	KO	411	3694	☆	0	0.962	Streptophyta	B9ICE5_POPTR Glutamate dehydrogenase
P00367	-	KO	558	9606	★	0	0.670	Chordata	DHE3_HUMAN Glutamate dehydrogenase 1
P49448	-	KO	558	9606	★	0	0.707	Chordata	DHE4_HUMAN Glutamate dehydrogenase 2
Q64HZ8	-	KO	558	9598	★	0	0.705	Chordata	DHE4_PANTR Glutamate dehydrogenase 2
P26443	-	KO	558	10090	★	0	0.712	Chordata	DHE3_MOUSE Glutamate dehydrogenase 1

Figura 16: Página *web* para visualização dos resultados SeedServer. Atalhos representados por números em vermelho na Figura - 1: Botão para obtenção de um arquivo tabulado contendo todos os dados do(s) agrupamento(s) formado(s); 2: Botão para obtenção detalhada do LCA; 3: Botão para obtenção das sequências em formato FASTA; 4: Botão para geração de relatório detalhado de taxonomia e presença/ausência de proteínas em grupos taxonômicos relacionados ao LCA.

O relatório detalhado oferece dados taxonômicos do LCA obtido, mesmo que ele não tenha uma classificação estabelecida na taxonomia de Lineu, assim como dados do grupo taxonômico de Lineu mais próximo. Os dados taxonômicos são: identificador numérico, nome e nível do grupo. O texto do resultado é apropriado para ser utilizado por programas.

O relatório taxonômico foi dividido em três partes, a saber: 1- Relatório das linhagens taxonômicas: oferece para cada proteína uma lista completa com os nomes disponíveis para os principais níveis taxonômicos, desde super-reino até espécie; 2- Relatório de contagem dos táxons: oferece uma contagem dos táxons presentes em cada um dos níveis taxonômicos; 3- Relatório para presença/ausência das proteínas em estudo em genomas pertencentes aos grupos taxonômicos relacionados ao LCA, com suporte da presença ou não de genomas completos nos respectivos grupos, sendo subdivido em: 3A- Mesmo nível taxonômico do LCA (grupos irmãos do LCA) e 3B- Um nível mais folha ao LCA (grupos filhos do LCA). A Figura 17 retrata um exemplo.

O fato de uma proteína não ter sido encontrada em um determinado táxon pode se justificar simplesmente pelo fato da falta de amostragem, dessa forma, evidenciar a presença de genomas completos em grupos irmãos do LCA ou grupos filhos, ajuda na inferência correta da origem do gene, bem como na dedução de quais grupos filhos do LCA herdaram o mesmo.

Para exemplificar a usabilidade desse relatório, um estudo do nosso grupo ainda em desenvolvimento, já mencionado na seção 6.1.5, referente à via de proteínas envolvidas em resistência a seca de plantas, teve todos seus grupos de homólogos inspecionados. Um quadro completo retratando a ausência de proteínas a partir de determinados grupos taxonômicos, com suporte de genomas completos, está mostrado na Figura 18. Os genes representados por círculos estão presentes em *A. thaliana* e ausentes no grupo irmão indicado, enquanto por retângulo, ausentes em todos os irmãos da linhagem de *A. thaliana*. Por exemplo, ABA2 não ocorre somente em archaea, mesmo com 122 genomas completos.

Um estudo ainda mais abrangente abordando deleções deduzidas dos grupos de homólogos será mostrado ao longo da próxima seção.

Same LCA Level Taxonomic Report for cluster 1:

LCA METAZOA AT THE KINGDOM LEVEL:

Rank Name	Complete Genomes	Assembly Genomes	Incomplete/Unfinished Genomes	No status Genomes
Viridiplantae	0/4	0/27	0/73	0/118051
Metazoa	3/4	42/137	8/136	25/267034
Fungi	0/14	0/180	0/153	0/69259

One LCA Level Down Taxonomic Report for cluster 1:

GENES FOUND BY SEEDSERVER AT THE PHYLUM LEVEL:

Rank Name	Complete Genomes	Assembly Genomes	Incomplete/Unfinished Genomes	No status Genomes
Placozoa	0/0	1/1	0/0	0/94
Cnidaria	0/0	1/3	0/3	0/4750
Nematoda	1/1	2/18	0/14	0/5120
Arthropoda	0/0	17/48	0/40	1/168585
Chordata	2/3	21/62	8/63	24/59332

GENES NOT FOUND BY SEEDSERVER AT THE PHYLUM LEVEL:

Rank Name	Complete Genomes	Assembly Genomes	Incomplete/Unfinished Genomes	No status Genomes
Porifera	0/0	0/1	0/0	0/1454
Ctenophora	0/0	0/0	0/1	0/81
Platyhelminthes	0/0	0/1	0/4	0/4992
Nemertea	0/0	0/0	0/0	0/443
Sipuncula	0/0	0/0	0/0	0/97
Mollusca	0/0	0/1	0/4	0/14464
Brachiopoda	0/0	0/0	0/0	0/218
Bryozoa	0/0	0/0	0/0	0/442
Priapulida	0/0	0/0	0/1	0/15

Figura 17: Página *web* contendo relatório da presença/ausência das proteínas em grupos taxonômicos no mesmo nível e um nível inferior ao LCA (filo *Metazoa*) que é mostrada com suporte da existência do total de genomas: completamente seqüenciados, em processo de montagem, incompletos ou sem projeto. Azul escuro: táxon onde proteína foi encontrada; Azul claro: táxon onde a proteína não foi encontrada. LCA: *Lowest Common Ancestor*.

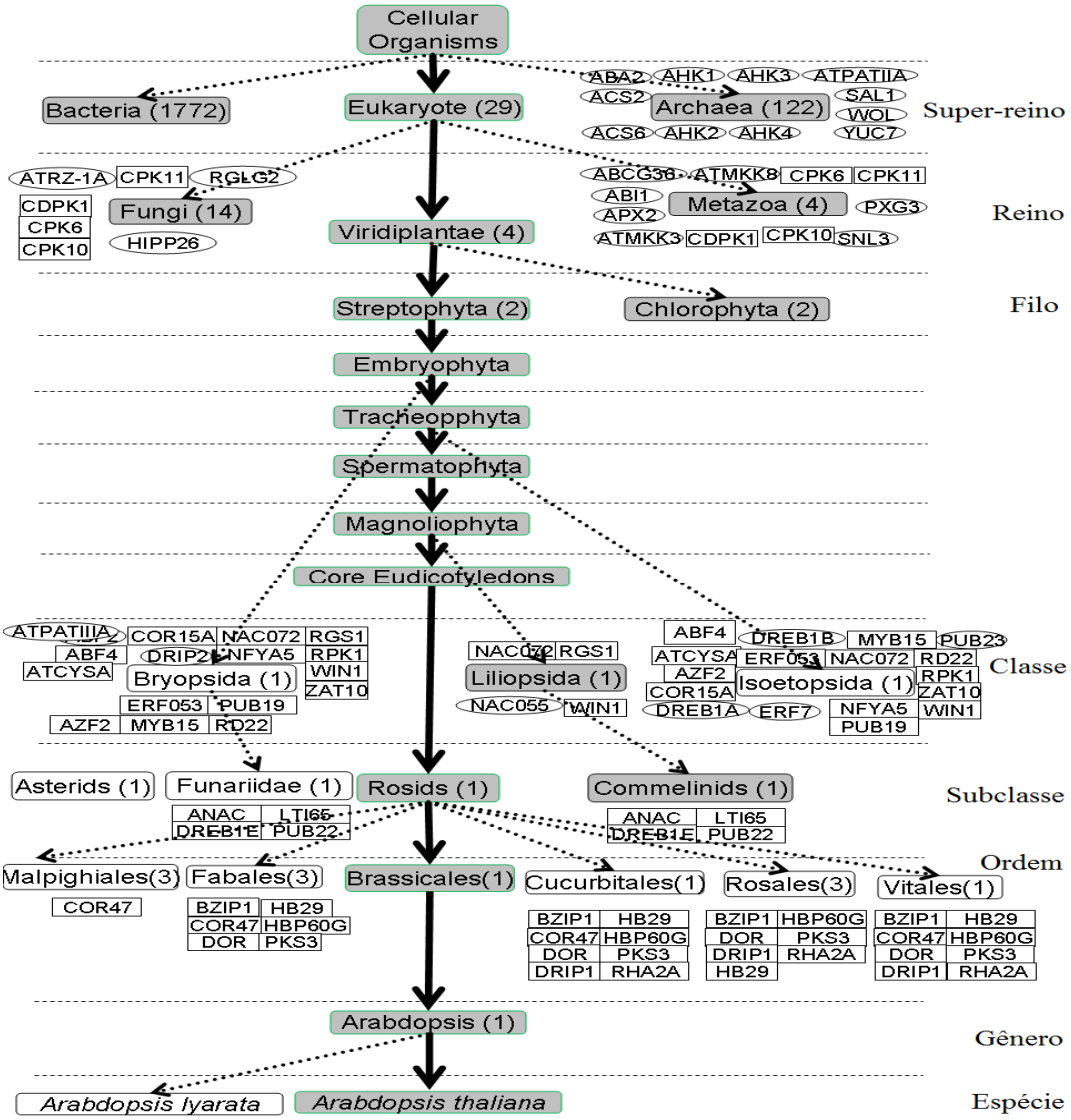


Figura 18: Ausência de proteínas envolvidas na resistência a seca em plantas em diversos grupos taxonômicos. Proteínas representadas por: Retângulos - estão ausentes em grupos taxonômicos de mesmo nível; Círculos – estão ausentes somente no respectivo grupo. Grupo taxonômico representados em: Cinza – possui genoma completo; Branco: possui genoma em montagem. Total de genomas completos ou em montagem mostrados entre parênteses. Grupos que possuem somente genomas incompletos não foram representados. Setas contínuas: taxonomia do organismo de referência *Arabidopsis thaliana.* Setas pontilhadas: taxonomia dos outros grupos representados. Níveis taxonômicos não intitulados não possuem classificação estabelecida (ex: *Embryophyta*).

6.2.SeedServer e o estudo da extinção de vias metabólicas

6.2.1. Aminoácidos essenciais

Grupos de homólogos foram criados para enzimas das vias biossintéticas dos aminoácidos essenciais (AEs) em humanos, através da metodologia SeedServer, tendo como objetivo o estudo de eventos de deleções gênicas. Para isso, foram utilizadas como *Seeds* enzimas de organismos sabidamente autotróficos, ou seja, capazes de sintetizá-los por vias *de novo*. Serina (S) e glicina (G), aminoácidos não-essenciais, estão presentes como controle. Como a busca reúne organismos de diversos níveis taxonômicos, que possuem ou não genomas completamente seqüenciados, os dados foram representados em nível de filo da seguinte forma, como visto na Figura 19: (a) círculos pretos para filos que possuem ao menos um genoma completo; (b) círculos cinza para filos que contem no máximo genomas em montagem, ou seja, sem genomas completos e (c) círculos vazios para filos com apenas genomas sem projeto de sequenciamento. No intuito de detectar e incluir nas análises proteínas modeladas incorretamente e sequências parciais, foi incluída uma busca BLAST adicional utilizando-se também as sequências marcadas como *fragment* na base UniProtKB (ver métodos, seção 5.2.2) e os alinhamentos com qualquer membro do grupo formado com cobertura acima de 50%, sendo esses dados representados por triângulos para todos os filos (mesmo código de cor que dos círculos). Essa busca foi importante para investigar a ocorrência das enzimas em esponja, pois o genoma disponível era recente e as sequências não estavam no UniProtKB.

As enzimas estão dispostas em ordem seqüencial anabólica da esquerda para direita, sendo que algumas vias estão representadas por vias alternativas, indicadas

por números, como a via da serina S(1): a partir de 3P-D-glicerato e S(2): a partir de piruvato.

Como esperado, as vias dos aminoácidos não-essenciais serina e glicina foram tidas como potencialmente presentes em diversos filos de eucariotos, com algumas exceções, a saber: (i) a via de serina esta ausente em *Rhodophyta* apesar do genoma completo de *Cyanidioschyzon merolae* estar presente, no entanto todas as vias estudadas mostraram-se incompletas nesse grupo. A busca adicional BLAST e inspeções manuais corroboram esse dado, mesmo tendo sido encontrado uma proteína parcial na via G1 de glicina; (ii) A biossíntese de serina parece ser ausente em *Apicomplexa*, um grupo que contém dois genomas completos de *Plasmodium,* sem as enzimas S1 e S4, similarmente a *Rhodophyta* os outros aminoácidos também se mostraram como AEs; (iii) Considerando os animais metazoários, falhamos em comprovar a não essencialidade de glicina para *Mollusca*, porém ainda não há genomas completos nesse grupo; (iv) Para *Microsporidia*, fungos intracelulares obrigatórios tendo *E. cuniculi* com genoma completo, foi comprovada a ausência quase total dos genes de biossíntese de aminoácido [46], sendo, portanto serina e glicina também AEs.

A Figura 19 mostra claramente as vias completas para fungos (*Ascomycota* e *Basidiomycota*) e plantas (*Chlorophyta* e *Streptophyta*). Nesse estudo, tentamos demonstrar quando e como uma provável Grande Deleção Genômica (GDG) ocorreu, ou seja, como se deu a perda de diversas enzimas responsáveis pela síntese dos aminoácidos, em diversos filos eucarióticos. As evidências mostram que já antes da origem dos metazoários (*Choanozoa*), a perda simultânea da capacidade de síntese dos nove AEs já se encontrava estabelecida. A esse fenômeno inicial, se seguem diversas perdas secundárias e independentes, fato evidenciado pela presença

de K14, M4 e M9 em *Chordata* e ausência delas em *Arthropoda*. Curiosamente, enzimas como VIL1 e M7 são mantidas em diversos grupos de metazoários apesar das respectivas vias estarem incompletas, o que pode ser explicado pela participação em vias secundárias. Particularmente, essas enzimas são muito importantes no catabolismo dos respectivos aminoácidos, e isso pode ser importante à luz da heterotrofia para nitrogênio, discutida adiante.

De fato, a GDG ocorre também em grupos não metazoários. Diferentes filos contendo genomas completos como *Rhodophyta*, *Euglenozoa* e *Apicomplexa* mostram padrões semelhantes ao fenótipo de essencialidade. Além disso, há evidências de vias incompletas para os fungos *Microsporidia* e *Neocallimastigomycota*, o que demonstra que a GDG ocorreu em eventos separados em pelo menos outros três grupos taxonomicamente separados, no ancestral comum a *Chromalveolates* e *Excavates*, no grupo *Fungi-Metazoa* e nas plantas *Archaeplastida*.

Similarmente a *Choanozoa*, grupos como *Heterokontophyta* e *Rhizaria* apresentam várias enzimas e até algumas vias completas. Em *Heterokontophyta*, somente a via de histidina (H) não está completa enquanto para *Rhizaria*, valina (V), isoleucina (I), lisina (K) e treonina (T) são vias potencialmente completas. O mesmo ocorre para metionina (M) em *Euglenozoa* e *Amoebozoa*. No entanto, quando mais genomas completos estiverem disponíveis, é possível que outras vias sejam revistas como completas em alguns grupos.

A perda secundária e gradual de genes parece já ter se estabilizado partir de *Placozoa*, *Porifera* e *Cnidaria*.

Como o primeiro genoma de *Porifera* ainda é um projeto em andamento e suas proteínas não se encontravam disponíveis no UniProtKB, inspecionamos

manualmente o proteoma deduzido usando alinhamentos BLAST filtrados, sendo observado o mesmo quadro de essencialidade para os nove aminoácidos.

É importante observar que grupos sem nenhuma proteína amostrada não foram representados, como *Apusozoa* e *Jakobida*.

Eukaryota
Cryptophyta
Haptophyta
Rhizaria
Heterokontophyta
Alveolata
Ciliophora
Apicomplexa
Euglenozoa
Excavates
Parabasalia
Fornicata
Amoebozoa
Unikonts
Neocallimastigomycota
Microsporidia
Eukarya
Basidiomycota
Ascomycota
Choanozoa
Metazoa
Porifera
Placozoa
Cnidaria
Nematoda
Platyhelminthes
Mollusca
Arthropoda
Chordata
Archaeplastida
Glaucophyta
Rhodophyta
Streptophyta
Viridiplantae
Chlorophyta
ATP phosphoribosyltransferase H1
phosphoribosyl-ATP pyrophosphohydrolase H2
Phosphoribosyl-AMP cyclohydrolase H3
5-proFAR isomerase H4
Glutamine amidotransferase H5
Imidazoleglycerol-phosphate dehydratase H6
Histidinol-phosphate aminotransferase H7
Histidinol phosphatase H8
Histidinol dehydrogenase H9
3-deoxy-7-phosphoheptulonate synthase FW1
3-dehydroquinate synthase FW2
3-dehydroquinate dehydratase FW3
Shikimate 5-dehydrogenase FW4
Shikimate kinase FW5
3-phosphoshikimate 1-carboxyvinyltransferase FW6
Chorismate synthase FW7
Chorismate mutase F8
Prephenate dehydratase F9
4 distinct aminotransferases F10
3-deoxy-7-phosphoheptulonate synthase FW1
3-dehydroquinate synthase FW2
3-dehydroquinate dehydratase FW3
Shikimate 5-dehydrogenase FW4
Shikimate kinase FW5
3-phosphoshikimate 1-carboxyvinyltransferase FW6
Chorismate synthase FW7
Anthranilate synthase W8
Anthranilate phosphoribosyltransferase W9
N(5`-phosphoribosyl) anthranilate isomerase W10
Indole-3-glycerol phosphate synthase W11
Tryptophan synthase W12
Acetolactate synthase VIL1
Ketol-acid reductoisomerase VIL2
Dihydroxy-acid dehydratase VIL3
Branched-chain amino acid aminotransferase VIL4
Threonine dehydratase I1
Acetolactate synthase VIL1
Ketol-acid reductoisomerase VIL2
Dihydroxy-acid dehydratase VIL3
Branched-chain amino acid aminotransferase VIL4
Pyruvate dehydrogenase L1
Acetolactate synthase VIL1
Ketol-acid reductoisomerase VIL2
Dihydroxy-acid dehydratase VIL3
2-isopropylmalate synthase L5
3-isopropylmalate dehydratase L6
3-isopropylmalate dehydrogenase L7
Branched-chain amino acid aminotransferase VIL4
Homocitrate synthase K1
Homoaconitate hydratase K2
Homoisocitrate dehydrogenase K3
Aromatic amino acid aminotransferase K4
Aminoadipate-semialdehyde dehydrogenase K5
Saccharopine dehydrogenase (NADP+, L-glutamate) K6
Saccharopine dehydrogenase (NAD+, L-lysine) K7
Aspartate kinase KMT8
Aspartate-semialdehyde dehydrogenase KMT9
Dihydrodipicolinate synthase K10
dihydrodipicolinate reductase K11
LL-diaminopimelate aminotransferase K12
Diaminopimelate epimerase K13
Diaminopimelate decarboxylase K14
Aspartate kinase KMT8
Aspartate-semialdehyde dehydrogenase KMT9
Homoserine dehydrogenase MT3
Homoserine O-acetyltransferase M4
Cystathionine gamma-synthase M5
Cystathionine beta-lyase M6
4 distinct Methyltransferases M7
Serine O-acetyltransferase M8
Cysteine synthase A M9
Cystathionine gamma-yase M10
Cystathionine beta-lyase M6
4 distinct Methyltransferases M7
Aspartate kinase KMT8
Aspartate-semialdehyde dehydrogenase KMT9
Homoserine dehydrogenase MT3
Homoserine kinase T4
Treonine synthase T5
D-3-phosphoglycerate dehydrogenase S1
Phosphoserine aminotransferase S2
Phosphoserine phosphatase S3
Serine/threonine dehydratase S4
3 distinct transaminases G1
Serine hydroxymethyltransferase G2
Threonine aldolase G3
H
F
W
V
I
L
K (1)
K (2)
M (1)
M (2)
T
S (1)
S (2)
G (1)
G (2)
G (3)

Figura 19: Representação esquemática para presença/ausência das enzimas biossintéticas dos nove aminoácidos essenciais e dos não essenciais serina e glicina. Código de uma letra utilizado para representar os aminoácidos, sendo H:histidina, F:fenilalanina, W:triptofano, V:valina, I:isoleucina, L:leucina, K:lisina, M:metionina, T:treonina, S:serina, G:glicina. Árvore taxonômica eucariótica amostrada em nível de filos. Círculos representam detecção pela metodologia SeedServer e triângulos pela busca BLAST adicional incluindo proteínas parciais. Preto: presença de ao menos um genoma completo; Cinza: presença de genomas em montagem mas não completos; Brancos: sem genomas completos ou em montagem. *Saccharomyces cerevisiae* (*Ascomycota*) e *Arabidopsis thaliana* (*Streptophyta*) foram usadas como *Seeds*. As 4 aminotransferases da via de fenilalanina são: (i) aspartato aminotransferase (ii) histidinol-fosfato aminotransferase (iii) aminotransferase de aminoácido aromático (iv) tirosina aminotransferase. As 4 metiltransferases na via de metionina são: (i) 5-metiltetrahidropteroiltriglutamato--homocisteina metiltransferase (ii) homocisteina S-metiltransferase (iii) betaina-homocisteina metiltransferase (iv) 5-metiltetrahidrofolato--homocisteina metiltransferase. As 3 transaminases na via de glicina são: alanina-glioxilato transaminase, serina-glioxilato transaminase e serina-piruvato transaminase.

6.2.2. Via biossintética de lisina

A via biossintética de lisina (K) é sabidamente distinta em plantas, fungos e archaeas. Nos fungos, está presente a via K(1) do α-aminoadipato [47] em oposição a K(2) do diaminopimalato, presente em plantas, algas e bactérias [48, 49]. Na Figura 20 inspecionamos o cenário completo da síntese de lisina apresentada na Figura 19, incluindo procariotos e uma terceira via K(3), comumente presente em archaeas, mas também encontrada completa em grupos bacterianos [50]. Por isso, *Seeds* de *Pyrococcus horikoshii* foram também usadas nessas análises.

Curiosamente, *Chlamydiae* parece representar um exemplo de essencialidade também entre procariotos uma vez que K14 se encontra ausente, genomas completos estão presentes e não se conhece enzima alternativa para essa reação. É possível observar também ausência de K12 em outros grupos bacterianos de genomas

completos (*Acidobacteria*, *Chlorobi*, *Deferribacteres*, *Deinococcus-Thermus*, *Fusobacteria*, *Chlamydiae*, *Synergistetes*, *Tenericutes* e *Thermotogae*), no entanto existem vias alternativas a essa enzima, que não foram abordadas nesse estudo, o que impossibilita inferir essencialidade.

A via K(3), como esperado, está completa em grupos de archaeas como *Crenarchaeota* e *Euryarcheota*, assim como nos grupos bacterianos *Chloroflexi* e *Proteobacteria*, sendo que possivelmente também em *Actinobacteria* e *Bacteroidetes*.

As primeiras quatro enzimas dessa via são coincidentes com K(1) - indicado por uma faixa cinza na Figura 20. A via K(1) está completa em *Proteobacteria*, fungos e possivelmente também em *Actinobacteria*, *Bacteroidetes* e *Firmicutes*, como evidenciado pela busca adicional BLAST por fragmentos (triângulos na Figura 20).

Grupos eucarióticos como *Rhizaria* e *Heterokontophyta* mantiveram a via K(2) provavelmente presente no ancestral dos organismos vivos, uma vez que a mesma se mostrou completa também em grupos de plantas, algas, bactérias e archaeas.

Em resumo, uma via presente em archaeas e em algumas bactérias (à direita na Figura 20), possivelmente dá origem à via de Lisina de fungos, pois as enzimas K1 a K4 são compartilhadas, enquanto a via anabólica de plantas é diferente (via central).

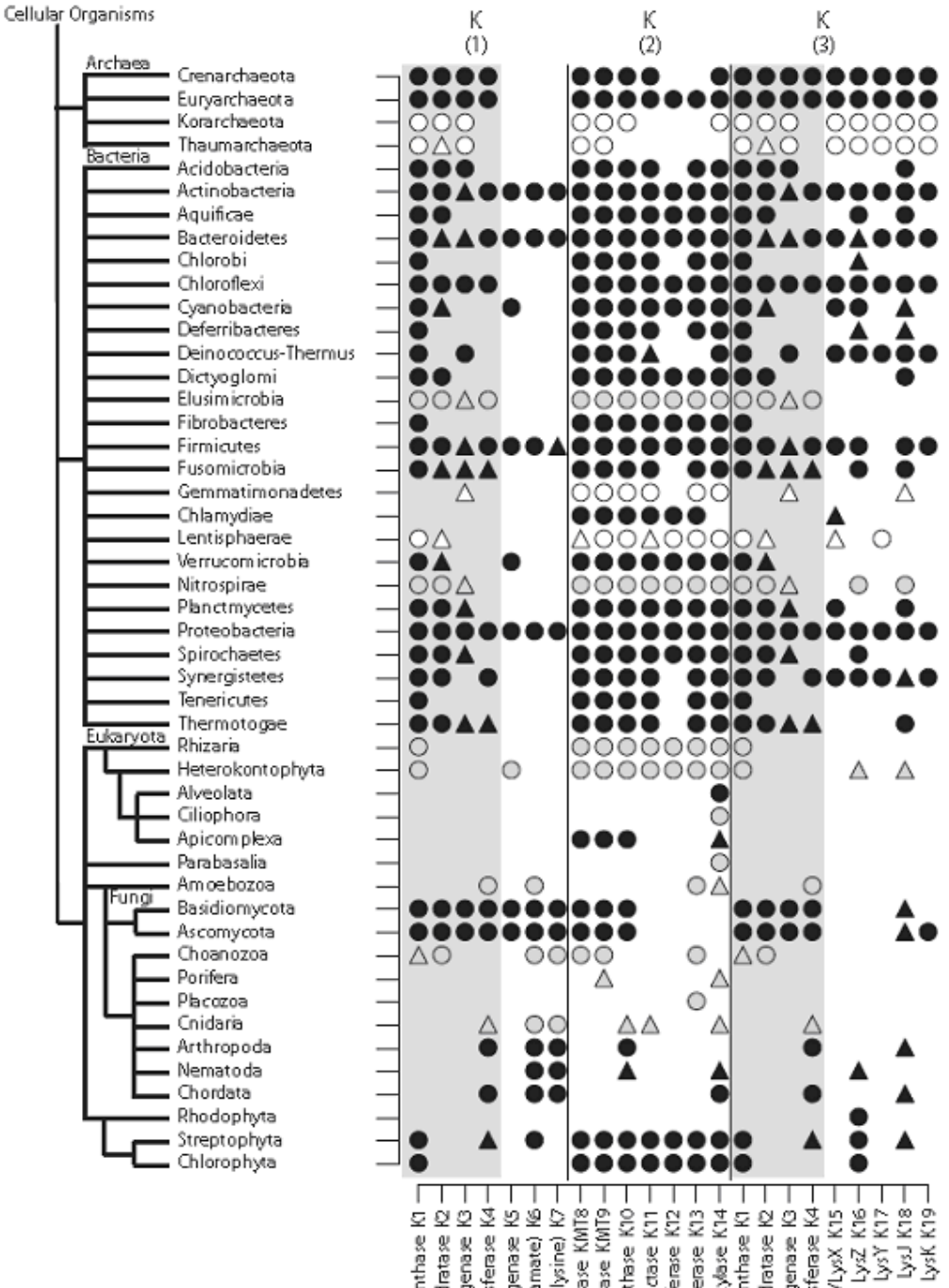

Cellular Organisms
Archaea
Crenarchaeota
Euryarchaeota
Korarchaeota
Thaumarchaeota
Bacteria
Acidobacteria
Actinobacteria
Aquificae
Bacteroidetes
Chlorobi
Chloroflexi
Cyanobacteria
Deferribacteres
Deinococcus-Thermus
Dictyoglomi
Elusimicrobia
Fibrobacteres
Firmicutes
Fusomicrobia
Gemmatimonadetes
Chlamydiae
Lentisphaerae
Verrucomicrobia
Nitrospirae
Planctmycetes
Proteobacteria
Spirochaetes
Synergistetes
Tenericutes
Thermotogae
Eukaryota
Rhizaria
Heterokontophyta
Alveolata
Ciliophora
Apicomplexa
Parabasalia
Amoebozoa
Fungi
Basidiomycota
Ascomycota
Choanozoa
Porifera
Placozoa
Cnidaria
Arthropoda
Nematoda
Chordata
Rhodophyta
Streptophyta
Chlorophyta
K (1)
K (2)
K (3)
Homocitrate synthase K1
Homoaconitate hydratase K2
Homoisocitrate dehydrogenase K3
Aromatic amino acid aminotransferase K4
Aminoadipate-semialdehyde dehydrogenase K5
Saccharopine dehydrogenase (NADP+, L-glutamate) K6
Saccharopine dehydrogenase (NAD+, L-lysine) K7
Aspartate kinase KMT8
Aspartate-semialdehyde dehydrogenase KMT9
Dihydrodipicolinate synthase K10
dihydrodipicolinate reductase K11
LL-diaminopimelate aminotransferase K12
Diaminopimelateepimerase K13
Diaminopimelate decarboxylase K14
Homocitrate synthase K1
Homoaconitate hydratase K2
Homoisocitrate dehydrogenase K3
Aromatic amino acid aminotransferase K4
LysW/LysX K15
LysZ K16
LysY K17
LysJ K18
LysK K19

Figura 20: Representação esquemática para presença/ausência das enzimas de biossíntese de lisina. K(1) representa a via de fungos (α-aminoadipato); K(2) a via de bactérias, algas e plantas (diaminopimelato); K(3) via de archaea, variante da via de α-aminoadipato. Árvore taxonômica amostrada em nível de filos. Círculos, triângulos e cores como explicados na Figura 19. *Saccharomyces cerevisiae* (*Ascomycota*), *Arabidopsis thaliana* (*Streptophyta*) e *Pyrococcus horikoshii* (*Euryarchaeota*) foram usados como *Seeds*.

6.2.3. Auxotrofia para nitrogênio

O consumo de aminoácidos é uma importante rota de assimilação de nitrogênio em outros compostos orgânicos para organismos heterotróficos como os cordados. A assimilação do amônio livre em eucariotos ocorre no citoplasma catalisado pela glutamato desidrogenase (EC:1.4.1.4) que incorpora o amônio em alfa-cetoglutarato gerando glutamato, usando elétrons do NADPH reduzido. Duas isoformas (denominadas assimilativas a partir daqui) dessa enzima estão presentes em fungos e uma em plantas, sendo que essas também possuem a opção de fixar nitrogênio com ajuda de bactérias nitrificantes.

A reação oposta é conduzida por outra isoforma mitocondrial da glutamato desidrogenase (EC:1.4.1.2) que carrega elétrons na coenzima NAD+ retirados do glutamato, causando sua decomposição e liberação de amônio.

Portanto, para investigar se a GDG das vias biossintéticas de AEs co-ocorreram com a perda da capacidade de assimilação de nitrogênio, criamos grupos de homólogos a partir das isoformas assimilativas de fungos, *Saccharomyces cerevisiae* (*GDH1* e *GDH3*) e, como controle, as isoformas mitocondriais (*GDH2*) catabólicas. Proteínas de *Arabidopsis thaliana* também foram usadas como *Seeds*: uma conhecida como *GDH* putativa que agrupou com as assimilativas de fungos e três

catabólicas, que agruparam com as mitocondriais humanas *GLUD1* e não com *GDH2* de fungos.

Os resultados estão na Figura 21A. A coluna da esquerda representa as isoformas assimilativas de fungos e plantas. As isoformas catabólicas mitocondriais de fungos (coluna central) e de plantas (coluna da direita) formaram grupos independentes.

Enzimas assimilativas foram encontradas em *Cnidaria*, grupo basal de metazoários, sendo todos os outros metazoários dependentes do consumo de aminoácidos para obtenção de nitrogênio e formação de compostos nitrogenados como DNA, incluindo até mesmo os poríferos. Apesar de fraca evidência devido à ausência de genomas completos, a assimilação também parece estar ausente em *Choanozoa* e *Placozoa*.

Comparando esses resultados aos da Figura 19, é notável que *Choanozoa*, tendo ainda tantas enzimas de biossíntese de aminoácidos, não seja capaz de assimilar nitrogênio. Aparentemente, a GDG se estabilizou em *Cnidaria*, sendo esse o último grupo metazoário com capacidade de assimilação.

A retenção de algumas enzimas mesmo quando a via se encontra incompleta, como já discutido, ocorre provavelmente por conexões em metabolismos secundários, como por exemplo: EC:1.2.1.31 aminoadipato-semialdeído desidrogenase K5 e EC:1.5.1.7 sacaropina desidrogenase K7, participam da via de degradação da lisina. À luz da dependência de nitrogênio que também se estabelece, é mais bem entendida a pressão seletiva para manutenção de algumas reações do catabolismo de AE.

Curiosamente, *GLUD1* de mamíferos possui um controle alostérico específico [51] que provavelmente regula a enzima para catalisar a reação no sentido de oxidar

o glutamato. Tal controle foi primeiramente observado em *Ciliophora* [52] cujo domínio responsável parece ter sido transferido por transferência gênica lateral a algum metazoário ancestral [53].

A Figura 21-B apresenta uma árvore filogenética com enzimas eucarióticas que sustentam o agrupamento separado dessas isoformas.

Outros grupos eucarióticos não metazoários como *Alveolata*, *Apicomplexa* e *Euglenozoa* possuem AEs (Figura 19), mas perderam a assimilação de nitrogênio (Figura 21-A) enquanto *Fornicata* e *Parabasalia* parecem ter mantido essa capacidade. Sem a detecção de nenhuma isoforma de glutamato desidrogenase se encontra *Rhizaria* (círculos cinza). Em *Rhodophyta* (círculos pretos) surpreendentemente nenhuma isoforma catabólica foi detectada.

Essas análises também foram conduzidas em procariotos. Formas assimilativas não foram detectadas em *Aquificae*, *Chlamydiae* e *Synergistetes*, todos contendo genomas completos. Essa ausência é condizente com a auxotrofia indicada por *Chlamydiae* na Figura 20 e dá suporte a idéia de que esse quadro pode ocorrer também em grupos procarióticos.

De uma forma geral, os dados mostram uma tendência à auxotrofia de nitrogênio sucedendo-se logo após o surgimento da essencialidade dos aminoácidos em metazoários, sendo *Cnidaria* uma exceção.

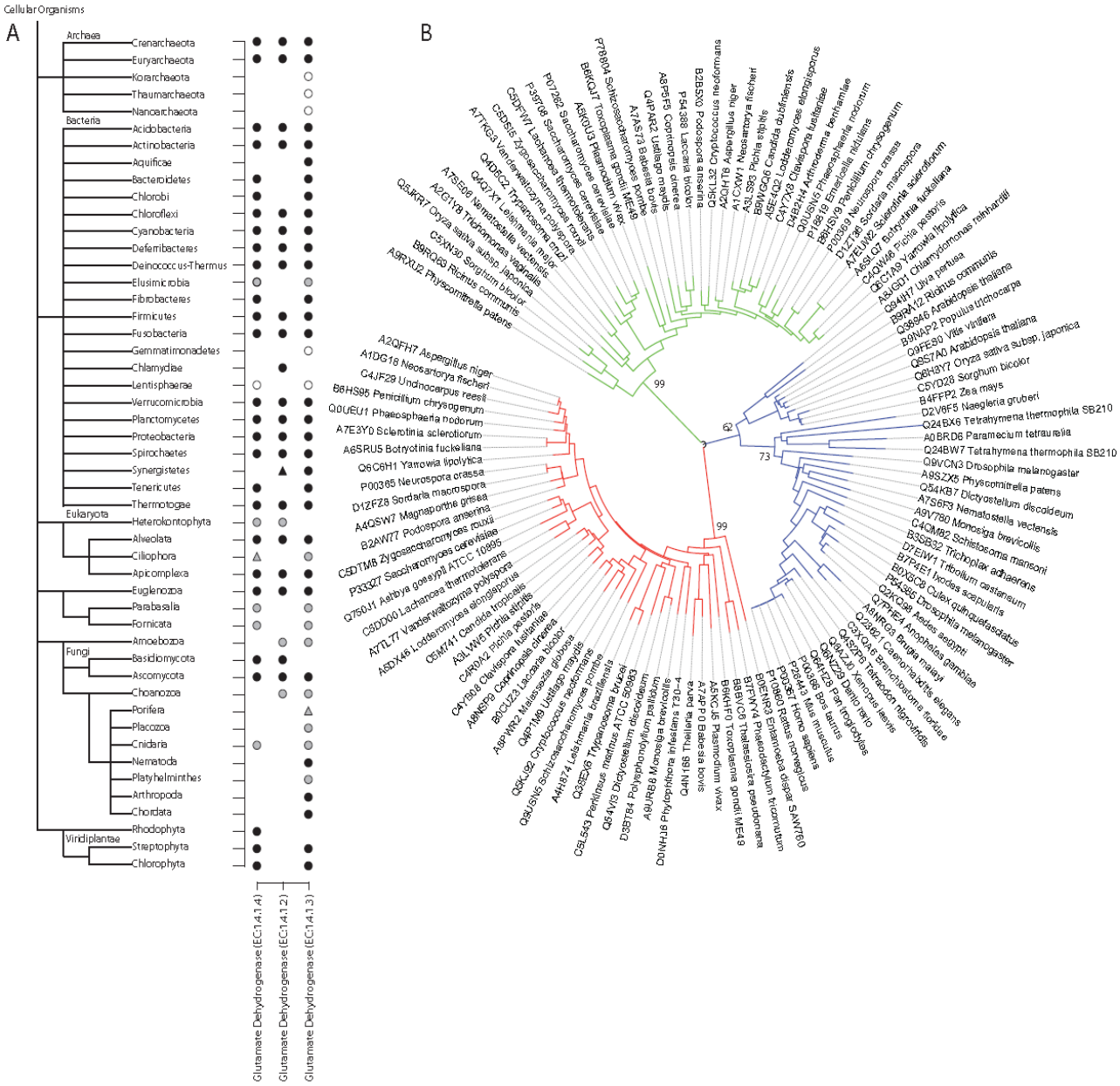


Figura 21: Representação esquemática para presença/ausência de glutamato desidrogenases. A: Coluna da esquerda: isoformas assimilativas *GDH1* e *GDH3* de *Saccharomyces cerevisiae* e *GDH* putativa de *Arabidopsis thaliana*; Coluna central: isoforma catabólica *GDH2* de *Saccharomyces cerevisiae;* Coluna da direita*:* isoformas catabólicas de *Arabidopsis thaliana.* Árvore taxonômica mostrada em nível de filo. Círculos, triângulos e cores como na Figura 19. *Saccharomyces cerevisiae* (*Ascomycota*) e *Arabidopsis thaliana* (*Streptophyta*) foram usadas como *Seeds*. B: Árvore filogenética com sequências eucarióticas das isoformas de glutamato desidrogenase. Ramos verdes: EC:1.4.1.4; Ramos vermelhos: EC:1.4.1.2; Ramos azuis: EC:1.4.1.3.

6.2.4. Enzimas remanescentes nas vias de AEs

Um estudo complementar com algumas enzimas remanescentes nas vias biossintéticas dos AEs (Figura 19) foi realizado para verificar se elas são mais susceptíveis a mutações quando comparadas a enzimas de vias funcionais. Também é possível que uma sub-funcionalização de parálogos tenha ocorrido em um ancestral comum de plantas, fungos e animais sendo que uma cópia divergente foi mantida em detrimento da original. Considerando ambas as hipóteses acima, comparamos filogeneticamente enzimas de aminoácidos essenciais e não essenciais que estivessem presentes em metazoários.

A Figura 22-A mostra uma árvore filogenética de acetolactato sintase (VIL1, Figura 19) e um grupo de alanina-glioxilato, serina-glioxilato e serina-piruvato transaminases (G1, Figura 22-B). Recordando, VIL1 é remanescente da biossíntese de AEs, enquanto G1 pertence a uma via completa em animais. Como esperado, a distância entre os ancestrais dos dois grupos prototróficos, *Streptophyta* (círculos verdes) e fungos (círculos amarelos) varia pouco: 0.4 e 0.7 para VIL1 e G1 respectivamente. A distância do ancestral de planta para o ancestral de *Metazoa* (círculos vermelhos) é relativamente maior na enzima remanescente VIL1: 1.0 (quando comparando com 0.4 representa aumento de 2,5 vezes) do que para G1: 0.7 (quando comparado com 0.7 não representa aumento). Portanto, essas enzimas presentes em vias incompletas parecem estar sofrendo maiores divergências evolutivas desde o surgimento da auxotrofia.

Para dar maior suporte a essa observação, a Figura 23 mostra razões de distância, mostradas no eixo Y esquerdo, calculadas entre os ancestrais para 12 enzimas.

O eixo Y direito corresponde à distância medida do ancestral de planta (*Streptophyta*) ao ancestral de fungo (*Dikarya*). Essa distância foi utilizada para normalizar as outras medidas de distância tomadas "De" (plantas: barras verdes, fungos: barras amarelas) "Para" outros grupos indicados no eixo X.

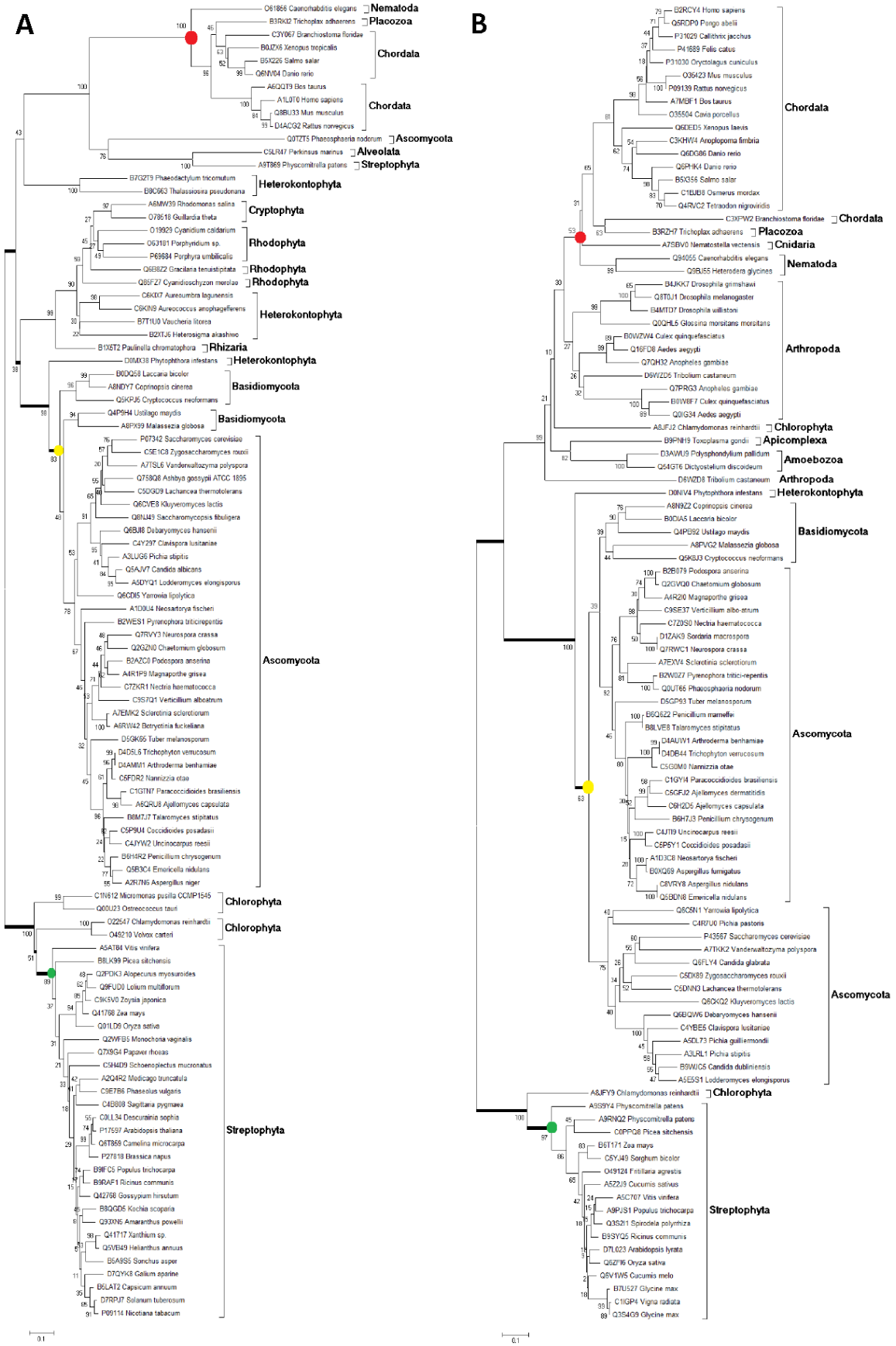

A
O61866 Caenorhabditis elegans
Nematoda
B3RKI2 Trichoplax adhaerens
Placozoa
C3Y067 Branchiostoma floridae
B0JZX6 Xenopus tropicalis
B5X226 Salmo salar
Q6NV04 Danio rerio
Chordata
A6QQT9 Bos taurus
A1L0T0 Homo sapiens
Q8BU33 Mus musculus
D4ACG2 Rattus norvegicus
Chordata
Q0TZT5 Phaeosphaeria nodorum
Ascomycota
C5LR47 Perkinsus marinus
Alveolata
A9T969 Physcomitrella patens
Streptophyta
B7G2T9 Phaeodactylum tricornutum
B8C963 Thalassiosira pseudonana
Heterokontophyta
A6MW39 Rhodomonas salina
O78618 Guillardia theta
Cryptophyta
O19929 Cyanidium caldarium
O63181 Porphyridium sp.
P69684 Porphyra umbilicalis
Rhodophyta
Q6B8Z2 Gracilaria tenuistipitata
Rhodophyta
Q85FZ7 Cyanidioschyzon merolae
Rhodophyta
C6KDX7 Aureoumbra lagunensis
C6KIN9 Aureococcus anophagefferens
B7T1U0 Vaucheria litorea
B2XTJ6 Heterosigma akashiwo
Heterokontophyta
B1X5T2 Paulinella chromatophora
Rhizaria
D0MX38 Phytophthora infestans
Heterokontophyta
B0DQ58 Laccaria bicolor
A8NDY7 Coprinopsis cinerea
Q5KPJ6 Cryptococcus neoformans
Basidiomycota
Q4P9H4 Ustilago maydis
A8PX99 Malassezia globosa
Basidiomycota
P07342 Saccharomyces cerevisiae
C5E1C8 Zygosaccharomyces rouxii
A7TSL6 Vanderwaltozyma polyspora
Q758Q8 Ashbya gossypii ATCC 1895
C5DGD9 Lachancea thermotolerans
Q6CVE8 Kluyveromyces lactis
Q8NJ49 Saccharomycopsis fibuligera
Q6BJI8 Debaryomyces hansenii
C4Y297 Clavispora lusitaniae
A3LUG6 Pichia stipitis
Q5AJV7 Candida albicans
A5DYQ1 Lodderomyces elongisporus
Q6CDI5 Yarrowia lipolytica
A1D0U4 Neosartorya fischeri
B2WES1 Pyrenophora triticirepentis
Q7RVY3 Neurospora crassa
Q2GZN0 Chaetomium globosum
B2AZC8 Podospora anserina
A4R1P9 Magnaporthe grisea
C7ZKR1 Nectria haematococca
C9S7Q1 Verticillium alboatrum
A7EMK2 Sclerotinia sclerotiorum
A6RW42 Botrytinia fuckeliana
D5GK66 Tuber melanosporum
D4DSL6 Trichophyton verrucosum
D4AMM1 Arthroderma benhamiae
C6FDR2 Nannizzia otae
C1GTN7 Paracoccidioides brasiliensis
A6QRU8 Ajellomyces capsulata
B6M7J7 Talaromyces stipitatus
C5P9U4 Coccidioides posadasii
C4JYW2 Uncinocarpus reesii
B6HMR2 Penicillium chrysogenum
Q5B3C4 Emericella nidulans
A2R7N6 Aspergillus niger
Ascomycota
C1N612 Micromonas pusilla CCMP1545
Chlorophyta
Q00U23 Ostreococcus tauri
Chlorophyta
O22547 Chlamydomonas reinhardtii
O49210 Volvox carteri
A5AT84 Vitis vinifera
B8LK99 Picea sitchensis
Q2PDK3 Alopecurus myosuroides
Q9FUD8 Lolium multiflorum
C9K5V0 Zoysia japonica
Q41768 Zea mays
Q01LD9 Oryza sativa
Q2WFB5 Monochoria vaginalis
Q7X9G4 Papaver rhoeas
C5H4D9 Schoenoplectus mucronatus
A2Q4R2 Medicago truncatula
C9E7B6 Phaseolus vulgaris
C4B888 Sagittaria pygmaea
C0LL34 Descurainia sophia
P17597 Arabidopsis thaliana
Q6T859 Camelina microcarpa
P27818 Brassica napus
B9IFC5 Populus trichocarpa
B9RAF1 Ricinus communis
Q42768 Gossypium hirsutum
B8QGD5 Kochia scoparia
Q93XN5 Amaranthus powellii
Q41717 Xanthium sp.
Q5VB49 Helianthus annuus
B5A9S5 Sonchus asper
D7QYK8 Gallium aparine
B6LAT2 Capsicum annuum
D7RPJ7 Solanum tuberosum
P09114 Nicotiana tabacum
Streptophyta
0.1
B
B2RCY4 Homo sapiens
Q6RDP0 Pongo abelii
P31029 Callithrix jacchus
P41689 Felis catus
P31030 Oryctolagus cuniculus
O35423 Mus musculus
P09139 Rattus norvegicus
A7MBF1 Bos taurus
O35504 Cavia porcellus
Q6DED5 Xenopus laevis
C3KHW4 Anoplopoma fimbria
Q6DG86 Danio rerio
Q6PHK4 Danio rerio
B5X356 Salmo salar
C1BUB8 Osmerus mordax
Q4RVC2 Tetraodon nigroviridis
Chordata
C3XPW2 Branchiostoma floridae
Chordata
B3RZH7 Trichoplax adhaerens
Placozoa
A7SBV0 Nematostella vectensis
Cnidaria
Q94055 Caenorhabditis elegans
Q9BJE5 Heterodera glycines
Nematoda
B4JKK7 Drosophila grimshawi
Q8T0J1 Drosophila melanogaster
B4MTD7 Drosophila willistoni
Q0QHL5 Glossina morsitans morsitans
B0WZW4 Culex quinquefasciatus
Q16FD8 Aedes aegypti
Q7QH32 Anopheles gambiae
D6WZD5 Tribolium castaneum
Q7PRG3 Anopheles gambiae
B0WBF7 Culex quinquefasciatus
Q0IG34 Aedes aegypti
Arthropoda
A8JFJ2 Chlamydomonas reinhardtii
Chlorophyta
B9PNH9 Toxoplasma gondii
Apicomplexa
D3AWU9 Polysphondylium pallidum
Q54GT6 Dictyostelium discoideum
Amoebozoa
D6WZD8 Tribolium castaneum
Arthropoda
D0NIV4 Phytophthora infestans
Heterokontophyta
A8N9Z2 Coprinopsis cinerea
B0DIA5 Laccaria bicolor
Q4PB92 Ustilago maydis
A8PVG2 Malassezia globosa
Q6K8J3 Cryptococcus neoformans
Basidiomycota
B2B879 Podospora anserina
Q2GVQ0 Chaetomium globosum
A4R2I0 Magnaporthe grisea
C9SE37 Verticillium albo-atrum
C7ZDS0 Nectria haematococca
D1ZAK9 Sordaria macrospora
Q7RWC1 Neurospora crassa
A7EXV4 Sclerotinia sclerotiorum
B2W0Z7 Pyrenophora tritici-repentis
Q0UT65 Phaeosphaeria nodorum
D5GP93 Tuber melanosporum
B6Q6Z2 Penicillium marneffei
B8LVE8 Talaromyces stipitatus
D4AUW1 Arthroderma benhamiae
D4DB44 Trichophyton verrucosum
C5G0M0 Nannizzia otae
C1GYI4 Paracoccidioides brasiliensis
C5GFJ2 Ajellomyces dermatitidis
C6H2D5 Ajellomyces capsulata
B6H7J3 Penicillium chrysogenum
C4JTI9 Uncinocarpus reesii
C5P5Y1 Coccidioides posadasii
A1D3C8 Neosartorya fischeri
B0XQ69 Aspergillus fumigatus
C8VRY8 Aspergillus nidulans
Q5BDN8 Emericella nidulans
Ascomycota
Q6C5N1 Yarrowia lipolytica
C4R7U0 Pichia pastoris
P43567 Saccharomyces cerevisiae
A7TKK2 Vanderwaltozyma polyspora
Q6FLY4 Candida glabrata
C5DK89 Zygosaccharomyces rouxii
C5DNN3 Lachancea thermotolerans
Q6CKQ2 Kluyveromyces lactis
Q6BQW6 Debaryomyces hansenii
C4YBE5 Clavispora lusitaniae
A5DL73 Pichia guilliermondii
A3LRL1 Pichia stipitis
B9WIC5 Candida dubliniensis
A5E5S1 Lodderomyces elongisporus
Ascomycota
A8JFY9 Chlamydomonas reinhardtii
Chlorophyta
A9S9Y4 Physcomitrella patens
A9RNQ2 Physcomitrella patens
C0PFQ8 Picea sitchensis
B6T171 Zea mays
C5YJ49 Sorghum bicolor
O49124 Fritillaria agrestis
A5Z2J9 Cucumis sativus
A5C767 Vitis vinifera
A9PJS1 Populus trichocarpa
Q3S2I1 Spirodela polyrhiza
B9SYQ5 Ricinus communis
D7L023 Arabidopsis lyrata
Q6ZFI6 Oryza sativa
Q6V1W5 Cucumis melo
B7U527 Glycine max
C1IGP4 Vigna radiata
Q3S4G9 Glycine max
Streptophyta
0.1

Figura 22: Árvores filogenéticas de A: acetolactato sintase (VIL1) enzima da via dos AEs valina, isoleucina e leucina e B: um grupo de alanina-glioxilato, serina-glioxilato e serina-piruvato transaminases (G1), enzimas da via de biossintese de glicina, aminoacido não essencial. Círculo vermelho: ancestral comum de metazoários; Círculo amarelo: ancestral comum de fungos; Círculo verde: ancestral de plantas *Streptophyta*. Em A: distância do círculo verde ao amarelo e vermelho são rescpectivamente 0.4 e 0.1. Em B: esses valores são respectivamente 0.7 e 0.7.

As enzimas S1, G1 e G2, pertencentes a vias de aminoácidos não essenciais, possuem razões baixas. Para as enzimas H5, FW7, F8, VIL1, VIL3, MT3 e M7 as razões mostradas pelas barras são alternativamente mais altas, chegando ao valor 7. A via M(2) está incompleta para *Basidiomycota*, porém MT3 está presente na via de treonina, que se encontra completa. Esses dados preliminares sugerem que modificações evolutivas ocorreram em diferentes níveis nas enzimas remanescentes de AEs.

K6 e K10 estão envolvidas em vias incompletas, respectivamente, de plantas e fungos. Correspondentemente, a distância medida de planta a fungo é alta, assim como a variação medida entre planta e *Chordata* (K6) e *Arthropoda* (K10), apresentando, portanto razões baixas. Como os ancestrais de fungos e plantas parecem igualmente distantes desses dois grupos e a divergência entre plantas e o grupo *Fungi*/*Metazoa* tende a uma trifurcação (ver Figura 22) as barras amarelas são similares às barras verdes, independentemente de quanta variação tenha ocorrido nas sequências do grupo animal.

Adicionalmente, uma análise cuidadosa das árvores filogenéticas parece indicar que uma sub-funcionalização de parálogos ocorreu em grupos basais como dos fungos e as cópias divergentes permaneceram nos grupos atuais de animais (Figuras suplementares disponíveis em [http://www.biodados.icb.ufmg.br/AEs/]).

É possível observar a presença de parálogos divergentes (out-parálogos) de *Streptophyta* e *Ascomycota* agrupados com sequências animais com *bootstrap* de 100% (Figura 22A). Similarmente, parálogos divergentes de *Dikarya* agrupam com sequências animais com *boostrap* de 98% para enzima M7 (Figura Suplementar 1) e para enzima K10 a ocorrência é entre sequências de *Streptophyta* e fungos, com *bootstrap* de 92% (Figura Suplementar 2).

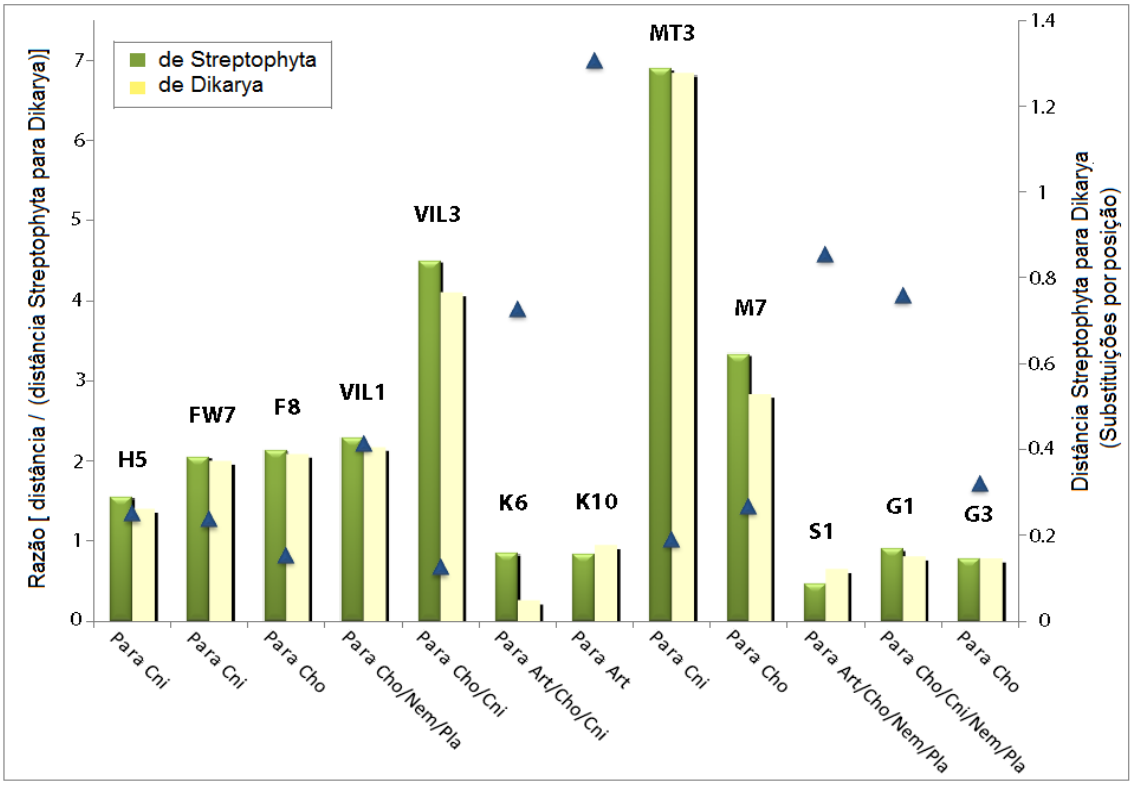


Figura 23: Distâncias relativas de enzimas das vias de aminoácidos essenciais e não essenciais de metazoários para homólogos presentes em fungos e plantas. Códigos para as enzimas são os mesmos para a Figura 19. No eixo Y direito, a distância entre o grupo *Streptophyta* de plantas e *Dikarya* de fungos foi utilizada para normalização e representada por triângulos. A distância "De" *Streptophyta* (barras verdes) e *Dikarya* (barras amarelas) "Para" grupos metazoário indicados no eixo X foi normalizada pela distância *Streptophyta*/*Dikarya* gerando a razão representada pelas barras no eixo Y esquerdo. S1, G1 e S2 são de vias de aminoácidos não essenciais. K6 e K10 são pertencentes a vias não completas em respectivamente *Streptophyta* e *Dikarya*. Abreviações: Art, *Arthropoda*; Cho, *Choanozoa*; Cni, *Cnidaria*; Nem, *Nematoda*; Pla, *Placozoa*.

Portanto, as enzimas remanescentes das vias de AEs mostram maiores divergências e isso pode ter ocorrido por neo-funcionalizações em grupos ancestrais.

6.3. Propagação de termos de ontologia funcional baseada em matrizes UniRef50

A propagação de informação de uma *Seed* para o grupo UniRef50 ao qual ela pertence não é feita sem um filtro de alinhamento dela com os membros do UniRef50 e avaliação da cobertura da *Seed* representada pelo alinhamento. Imaginamos um cenário onde matrizes de comparação par a par intra UniRef50 permitiriam o enriquecimento de qualquer base de dados, de maneira idêntica ao que é feito no UEKO e também no protocolo UE-*Seed*. Enquanto não disponibilizamos de suporte apropriado em supercomputadores, a produção dessas matrizes para todos os grupos UniRef50 permanece como uma meta. Todavia, conseguimos programar a criação dessas matrizes para todos os grupos UniRef50 de um conjunto restrito de organismos. Escolhemos para isso os organismos atualmente constituintes da base PANTHER. Para reduzir o custo computacional, um procedimento simples selecionou somente grupos UniRef50 que contém sequências de pelo menos um desses organismos e conduzimos alinhamentos com BLAST de todas contra todas, guardando as coberturas em matrizes.

Um total de 362.576 matrizes foram produzidas contendo 2.855.817 proteínas de 645 organismos, uma vez que outras sequências presentes em um determinado UniRef50 que não sejam de uma dos 82 organismos PANTHER também são consideradas. Essas matrizes são utilizadas em uma ferramenta que desenvolvemos visando propagar os termos GO (de ontologia gênica) associados a uma sequência para aquelas que satisfazem um limiar de cobertura desejado, 80% por exemplo.

Como aquelas relações que passam pelo filtro são consideradas fortes, tanto termos de uma proteína podem ser propagados para aquelas do grupo UniRef50 que passam pelo filtro, como podem também ser adquiridos delas. Essa aplicação visa remediar o fato de que termos GO são frequentemente atribuídos a uma lista de genes de um organismo, por especialistas nele, e nem sempre os termos atribuídos a um gene são propagados para possíveis homólogos. Assim, alguns organismos têm ontologias bem especificadas, com vários níveis, enquanto a proteína ortóloga ou não tem anotação alguma, ou tem termos GO mais genéricos associados a ela.

Analisando as relações estabelecidas nas matrizes sob diferentes limiares de cobertura, 100%, 80% e 50%, houve aumento da porcentagem de proteínas com termos GO associados para quase todos os organismos PANTHER, sendo entre 0% e 19% no limite de cobertura de 100% e de 0% até 23% com cobertura de 50% (Figura 24-A). *Canis familiaris* apresentou o maior ganho, provavelmente pela proximidade filogenética de organismos modelo mamíferos enquanto *Schizosaccharomyces pombe*, por já estar bem anotado em relação aos seus organismos mais próximos, não apresentou nenhum ganho significativo.

Ainda que a melhora na quantidade de termos tenha sido baixo para muitas espécies (mesmo considerando o cumulativo dos três níveis de recobrimento U-MAGE analisados), o mesmo não se aplica à qualidade da informação funcional propagada. Uma vez que a estruturação do termo ontológico é hierárquica, quanto mais "folha", mais detalhada está uma determinada função e em contrapartida, quanto mais "raiz", mais generalizada. Dessa forma, foi amostrada a porcentagem de termos GO que foram mais especificados após o uso da ferramenta U-MAGE (Figura 24-B), sendo que essa representatividade foi de até 54% em *Pan troglodytes* para 100% de cobertura.

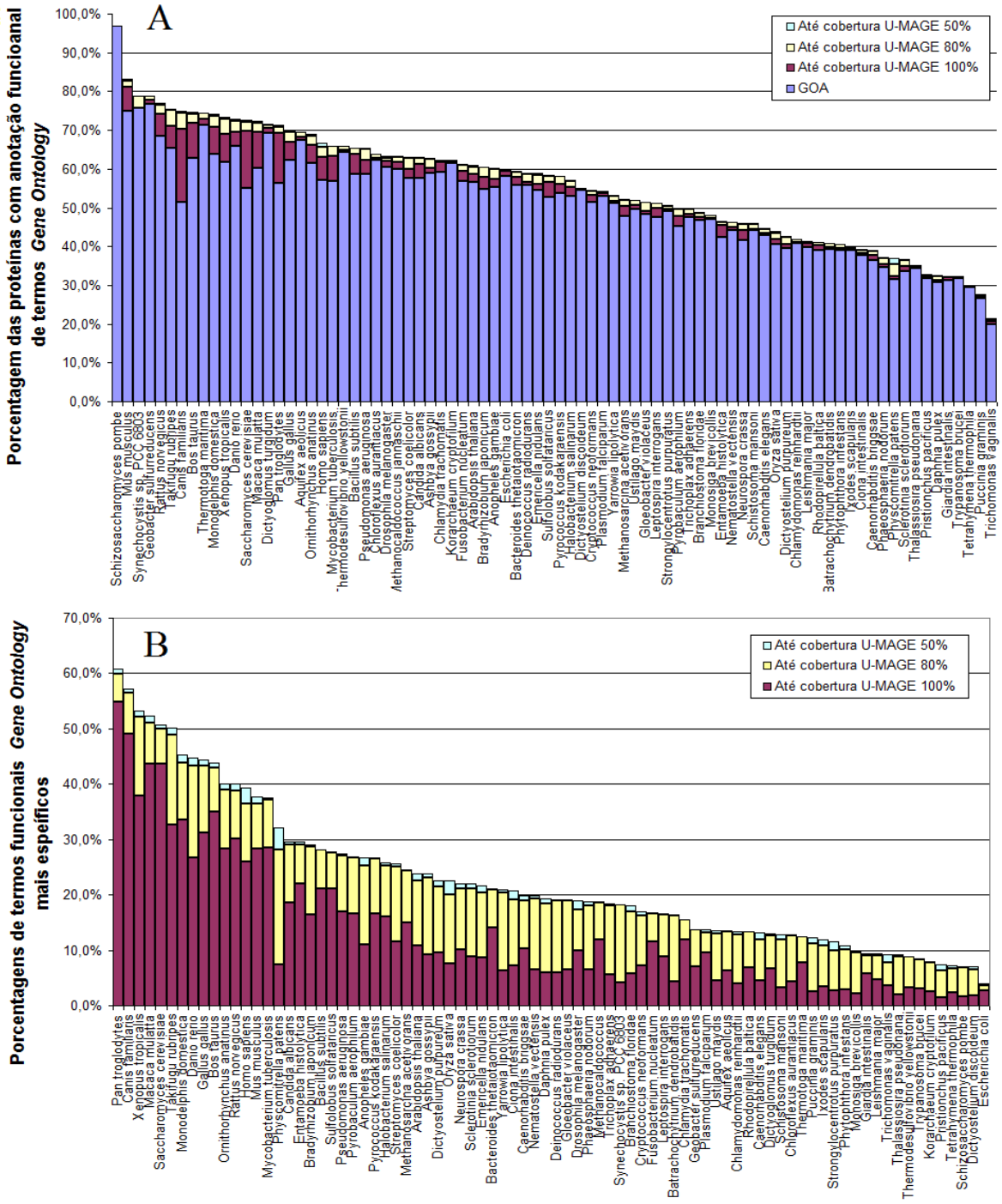


Figura 24: Enriquecimento quantitativo e qualitativo da propagação U-MAGE para 82 organismos da base de dados PANTHER. A: Enriquecimento adicional cumulativo por organismo de termos *Gene Ontology* associados a proteínas presentes em matrizes. B: Enriquecimento adicional cumulativo por organismo de termos *Gene Ontology* mais específicos associados a proteínas presentes em matrizes.

Esse efeito é ainda maior se o limiar de cobertura escolhido for menor. É importante notar que se a proteína não participa de nenhum grupo UniRef50, não participa das matrizes e não recebe propagação de termos GO.

6.3.1. U-MAGE em interface *web*

Novamente, uma interface *web* [http://pinguim.fmrp.usp.br/u-mage/form_u-mage.html] foi desenvolvida para facilitar e incentivar o uso da ferramenta. Não é necessário o cadastramento de usuários uma vez que os resultados são disponibilizados no momento da consulta, sendo requeridos como parâmetros apenas um identificador UniProtKB e um limite mínimo de cobertura desejado (Figura 25). Também estão disponíveis um guia passo a passo, explicando fundamentos e interpretação dos resultados, arquivos contendo todas as matrizes e novas relações entre identificadores UniProtKB e GO e por fim atalhos redirecionando o usuário para páginas externas.

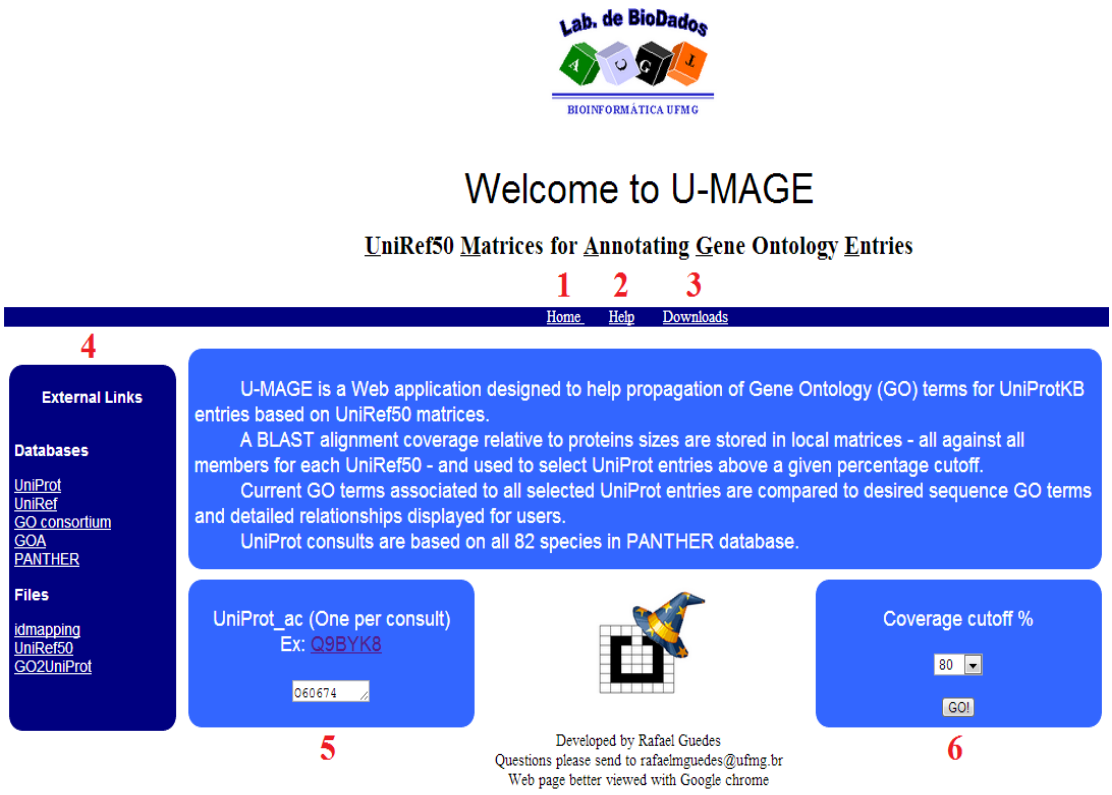


Figura 25: Página *web* principal desenvolvida para disponibilização do U-MAGE. Atalhos representados por números em vermelho na Figura - 1: Página principal para escolha dos parâmetros; 2: Guia de ajuda; 3: Página para obtenção de arquivos; 4: Atalhos para arquivos e páginas externas; 5: Campo para indicar o identificador UniProtKB a ser consultado; 6: Campo para indicar o limite de cobertura a ser utilizado na consulta.

Depois de efetuada uma consulta, a matriz UniRef50 correspondente é verificada para inspeção dos valores de cobertura e uma nova página contendo os resultados da propagação irá aparecer, organizada em forma de tabelas. Quando a cobertura da proteína sob consulta é superior ao limite dado com outras proteínas da matriz e diferenças entre os termos GO associados são encontradas, a primeira tabela de resultados é gerada.

As colunas representam todos os termos GO que são encontrados associados à proteína enviada como consulta. Esses termos são ordenados de acordo com a

hierarquia ontológica a que pertencem e cada hierarquia pode ser visualizada por uma cor distinta. Uma barra superior horizontal na cor azul é mostrada sempre que um termo listado é o mais específico da referida hierarquia para melhor identificação pelo usuário. Por sua vez, uma barra inferior horizontal da cor verde é mostrada para indicar que o código de evidência associado ao respectivo termo é de curadoria manual, pois nesses casos a propagação seria mais segura (Rachael Huntley, GOA, EBI, comunicação pessoal).

As linhas da primeira tabela são preenchidas com identificadores UniProtKB selecionados, para os quais os respectivos termos GO estejam-lhe ausentes.

De forma inversa, a segunda tabela de resultado relaciona proteínas selecionadas da matriz, cuja cobertura com a proteína sob consulta seja superior ao limite dado, mostrando quais termos estão ausentes na proteína enviada como consulta, mas presente em outras suficientemente similares do mesmo UniRef50.

Mais uma vez, barras horizontais azuis e verdes são utilizadas pra indicar especificidade do termo (mais "folha" na hierarquia) e curadoria manual, respectivamente. No entanto, como a informação de curadoria está associada a cada UniProtKB, essa marcação aparece em cada linha, e não no topo da coluna, como ocorre na primeira tabela.

Para ambas as tabelas, informações adicionais de identificação taxonômica e descrição dos termos GO e UniProtKB são apresentadas ao posicionar o *mouse* acima do termo desejado.

Adicionalmente, para indicar termos GO mais parentais que poderiam ser eliminados, uma vez que outros mais informativos estão sendo sugeridos, um identificar 'P' (de *Parent* em inglês) é mostrado juntamente com o identificador UniProtKB ou GO e a lista completa deles apresentada com o posicionamento do

mouse. De forma análoga, a letra 'C' (de *child* em inglês) é utilizada para evitar a propagação de termos mais parentais, uma vez que outros mais específicos (*children*) já estão presentes, também visualizados com o posicionamento correto do *mouse*.

Sempre que a geração de uma das tabelas não seja possível por falta de dados associados às proteínas, um aviso é apresentado na página. Identificadores UniProtKB que não apresentam nenhum termo GO ou cujos termos sejam todos idênticos à proteína sob consulta, são devidamente listados.

Para exemplificar o uso da ferramenta, a Figura 26 apresenta as duas tabelas de resultado para a proteína humana manualmente curada (SwissProt) tirosina quinase JAK2 (UniProtKB: O60674; EC: 2.7.10.2). Os termos mais específicos são destacados em cinza nas respectivas hierarquias funcionais e podemos constatar em ambas os quadros (Figuras 26 A e B) que mesmo proteínas manualmente curadas de organismos modelo podem estar com anotação funcional incompleta. Sugerimos que a proteína sob consulta O60674 deveria conter os termos GO:0043560, GO:0031702, GO:0051428, GO:0033130, GO:0008022 e GO:0043548 e perder termos mais parentais como GO:0005102. Ações similares ocorreriam com as proteínas selecionadas. A ferramenta disponibiliza na página para o usuário uma relação completa de ações de propagação de termos filhos e eliminação de termos parentais em um relatório que pode ser descarregado.

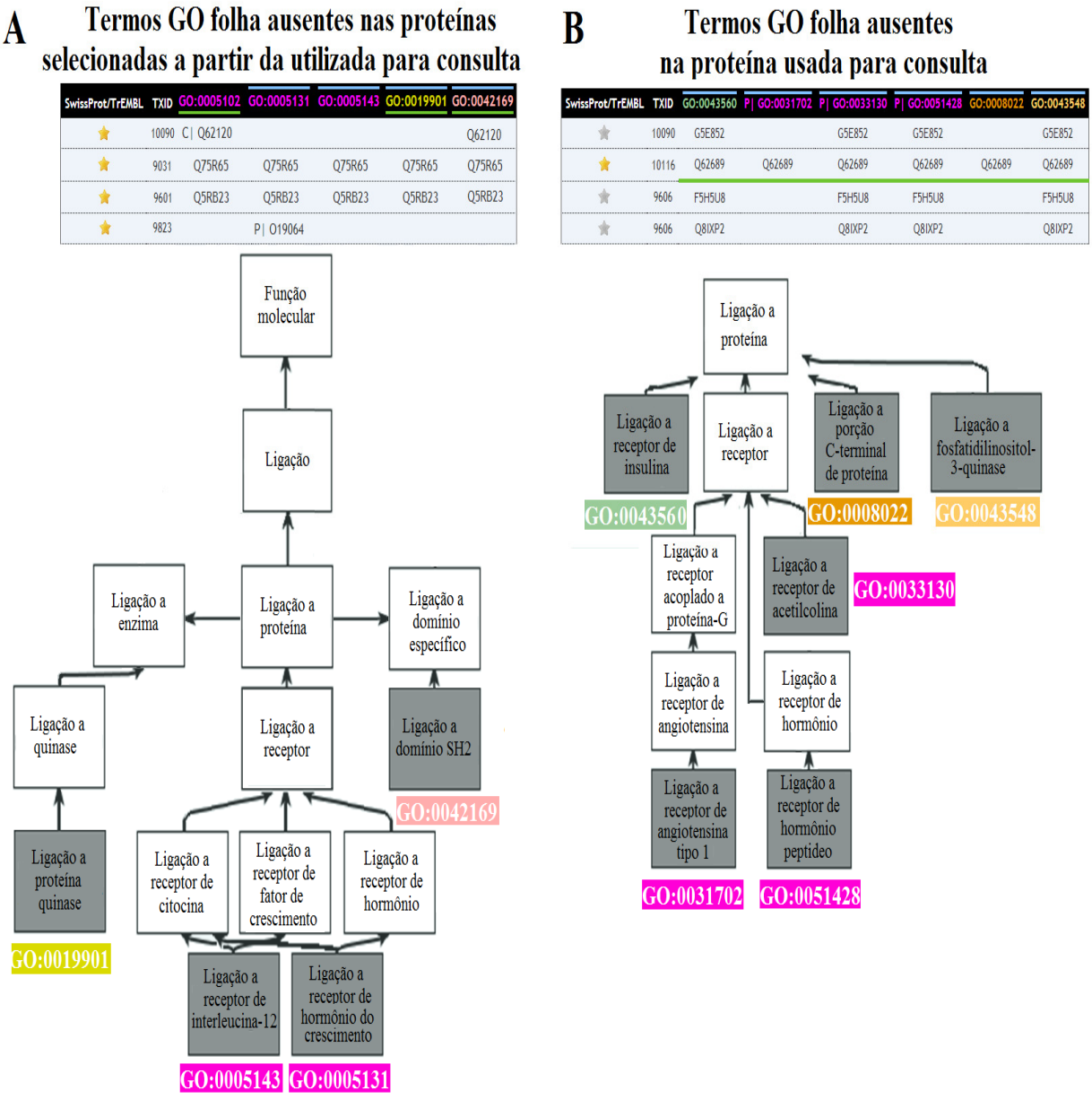


Figura 26: Propagação U-MAGE para a proteína humana O60674 (tirosina quinase JAK2, EC:2.7.10.2). A: Termos GO mais específicos ausentes nas proteínas selecionadas acima da cobertura de 80% com a proteína sob consulta. B: Termos GO mais específicos ausentes na proteína sob consulta, acima da cobertura de 80% com as proteínas selecionadas. Hierarquias funcionais são mostradas abaixo das tabelas com os termos respectivamente listados. Termos mais específicos destacados em cinza. As cores representam mesma hierarquia direta. TXID: identificador taxonômico; GO: identificador *Gene Ontology*; P: termos GO parentais de mesma hierarquia presentes; C: termos GO mais específicos (filhos) de mesma hierarquia presentes.

7. Discussão

Na última versão do UniProtKB utilizada no presente trabalho de fevereiro de 2012, havia o surpreendente número de aproximadamente 17 milhões de sequências completas de diversos genomas. Enquanto isso, bancos de dados de homólogos como COG [4], KEGG *Orthology* [17] e OMA [7, 54] disponibilizam uma porção desse universo protéico contida no genoma completo de algumas centenas de organismos. SeedServer é um ferramenta desenvolvida com o intuito de abranger todas as sequências completas depositadas no UniProtKB em uma busca direcionada ao interesse do usuário.

A ferramenta possui um programa principal responsável por coordenar seus diversos módulos constituintes. Os módulos Seed Linkage e UEKO demonstraram a importante função de incluir com eficiência sequências provenientes de genomas incompletos ou em processo de montagem em todos os exemplos estudados, chegando a uma contribuição superior a 50% do total de proteínas. No entanto, mesmo com os exigentes parâmetros de recrutamento, a inclusão de alguns poucos candidatos espúrios pode ainda ser observada. Tais inclusões puderam ser detectadas e filtradas do recrutamento SeedServer final através do módulo de validação escolhido, baseado em matrizes PSI-BLAST, cuja composição é controlada pela própria sequência *Seed* de interesse fundadora do agrupamento.

A correlação entre validação PSI-BLAST e a nomenclatura EC esperada foi aferida, demonstrando uma alta taxa de acerto entre sequências SwissProt curadas manualmente, de 99,72%. Embora não esperado, a mesma proporção de verdadeiros positivos foi encontrada utilizando-se além das entradas SwissProt, as entradas que receberam anotação automática na base TrEMBL. Adicionalmente, uma segunda correlação com famílias PANTHER foi utilizada. Mais uma vez, foi observado um

bom desempenho no nível de família PANTHER, onde 86,94% das proteínas associadas aos termos que possuíam descrições especificadas estavam de acordo com o esperado para cada caso. Já para o nível mais específico de subfamílias, esse mesmo índice foi de 91,04%. A alta similaridade entre proteínas que desempenham funções diferentes é um aspecto que inevitavelmente leva ao erro, entretanto a similaridade elevada possivelmente denota origem comum e proximidade evolutiva.

Também é possível haver manutenção da função biológica com alta similaridade da estrutura secundária ainda que haja uma baixa conservação da sequência primária por acúmulo de mutações a partir de um ancestral comum. No entanto, esses casos fogem ao objetivo desse trabalho que utiliza métricas de comparação de estrutura secundária (SOV) como parâmetro complementar à inferência de homologia e não como princípio para tal finalidade, como realizado por outros métodos [55] e organizado por bases de dados como a SCOP [56]. Assim, depois de utilizar em vários estudos de caso o parâmetro SOV, preferimos não recomendar um limiar de corte para determinar homologia, mas preservamos a sua exposição no relatório de resultados por ser uma forma de inferir similaridade estrutural. A presença e a manutenção de elementos estruturais como alfa-hélices ou folhas beta são contra-intuitivamente naturais em sequências randômicas de aminoácidos [57], no entanto parece improvável que a conservação da estrutura como um todo e a ordem dos elementos, refletida no parâmetro SOV seja por mero acaso. Assim, valores altos são uma informação adicional relevante.

Ao analisarmos milhares de comparações estruturais provenientes dos diversos experimentos realizados, porcentagens de aproximadamente 70% de valores SOV foram observadas entre homólogos separados nos três diferentes super-reinos da vida (eucariotos, procariotos e archaeas) enquanto entre homólogos de mesmo

gênero essa taxa foi de aproximadamente 90%. No entanto, estudos com outros programas para prever estrutura secundária, ou com dados extraídos de estruturas resolvidas, são necessários para comprovação desses dados preliminares. Uma atenção especial deve ser dada nesse tipo de estudo para proteínas multifuncionais bem como entre homólogos de organismos multiplóides ou com diversos parálogos como as plantas. Esta evidência de menor similaridade estrutural entre parálogos sustenta a teoria denominada "Conjectura do Ortólogo" [58] onde ortólogos são mais conservados que parálogos, um conceito refutado por Nehrt *et al.* [59], mas cada vez mais comprovado experimentalmente [60–63].

Uma interface *web* foi desenvolvida para utilização do SeedServer oferecendo um serviço estável para processos de poucas dezenas de *Seeds*, sendo recomendada instalação local para trabalhos mais abrangentes através do guia de instalação fornecido. A limitação de *Seeds* pertencentes à base de dados UniProtKB sendo negado o uso de sequências próprias dos usuários se justifica pela inviabilidade de conferência dos dados como atribuição de taxonomia adequada. Esse aspecto, no entanto não se mostrou uma limitação nos diversos exemplos apresentados nesse trabalho. A ferramenta conta com o uso do serviço computacional baseado em *web services* BOWS, que permitirá a transposição do processamento SeedServer entre diferentes servidores com capacidade de alto processamento, mantendo a estrutura básica *web* criada.

O Seedserver mostrou-se uma plataforma eficaz para estudos de inferência do surgimento de genes e vias através da determinação do LCA [45] bem como de deleções gênicas. Nosso grupo mostrou em um estudo de caso um quadro atualizado de uma Grande Deleção Genômica de diversas enzimas envolvidas na biossíntese de alguns aminoácidos em grupos de eucariotos e procariotos [64], inicialmente

discutido para dez eucariotos de genomas completos por Payne e Loomis [31]. A esta deleção se segue a perda da capacidade assimiladora de nitrogênio, componente essencial na formação dos aminoácidos, demonstrando que genes com funções supérfluas em um metabolismo tendem ser eliminados na evolução, fato evidenciado pela compactação dos genomas de organismos parasíticos. Em contrapartida, proteínas mantidas em um genoma pertencentes a vias funcionalmente incompletas demonstraram uma maior susceptibilidade ao acúmulo de mutações o que pode levar a sub ou neo-funcionalização das mesmas. A inclusão controlada de sequências fragmentadas nesse estudo mostrou potencial para agregar informação, sendo uma meta a atualização da ferramenta SeedServer para lidar com as mesmas. Todavia, as sequências agrupadas podem ser prontamente descarregáveis pelo usuário e utilizadas para esse fim por método análogo ao utilizado por nosso grupo.

Uma vez que uma das funções de um grupo de homólogos é a inferência da função de genes desconhecidos associados a outros de função descrita, quanto maior o número de sequências anotadas, melhor será a qualidade desse processo. Um exemplo recente mostra a inferência de possíveis patógenos de espécies do gênero *Candida* através da propagação de termos funcionais de *Candida albicans* [65]. Hoje em dia, a atribuição de termos *Gene Ontology* (GO) por grupos como o *Gene Ontology Annotation* (GOA) através de métodos manuais ou computacionais é uma importante fonte de anotação de propriedades biológicas às proteínas. No entanto, a representatividade desses bancos de dados é enviesada para alguns grupos de organismos modelo e funções escolhidas previamente [25]. No sentido de expandir essa anotação para organismos próximos aos anotados atualmente foi criada a ferramenta U-MAGE baseada em matrizes de cobertura de grupos UniRef50. A

metodologia atual de criação de grupos UniRef50 exige uma cobertura de 80% entre recrutadas e proteína representativa do grupo, porém esse filtro não elimina a presença de sequências com menos da metade do tamanho da representativa quando são advindas de grupos UniRef100, fato que justifica o custo operacional de geração das matrizes. Inicialmente somente termos de uma das três hierarquias presentes no GO foram escolhidos, os de "Função Molecular", por apresentarem alta conservação entre homólogos [60], porém futuramente as outras hierarquias como Componente Celular também poderão ser incluídas. A melhoria quantitativa (número de proteínas que adquirem termos GO) e qualitativa (número de termos GO mais específicos na hierarquia que foram adicionados) poderá auxiliar em estudos de caracterização das funções presentes em determinadas amostragens com maior suporte obtido por métodos estatísticos [24, 66]. A propagação automática de termos menos específicos das hierarquias GO entre homólogos já demonstrou alta eficácia em um experimento com quatro eucariotos [67] sendo que a ferramenta U-MAGE poderá servir como uma plataforma para propagação manual de termos GO mais específicos e uma vez estabelecidos parâmetros e limites de qualidade, subsidiar uma melhor propagação automática. Essa automação já demonstrou rivalizar em qualidade com anotações manuais [68].

Uma vez que o U-MAGE se restringe às sequências homólogas altamente similares, futuramente é possível uma interação com o SeedServer para expandir os limites da propagação a homólogos mais distantes. Ambas as ferramentas desenvolvidas nesse trabalho contribuem na propagação da informação biológica disponível para proteínas conhecidas, representando uma contribuição à Bioinformática.

8. Considerações finais

Nesse trabalho objetivamos a integração de ferramentas utilizadas no estudo de sequências protéicas homólogas de forma confiável e de fácil usabilidade a todos os tipos de usuários, criando uma plataforma denominada SeedServer, fornecendo subsídios para estudos complementares de filogenia, origem de genes e vias metabólicas, duplicações e deleções gênicas. Um estudo de caso foi apresentado criando um cenário atualizado no estudo das vias biossintéticas de aminoácidos essenciais e a distribuição da presença ou ausência das respectivas enzimas em diversos grupos eucarióticos e procarióticos. Contemplamos também a melhoria na qualidade da informação biológica associada a essas sequências constituintes dos bancos de dados, através da propagação segura da anotação ontológica funcional a partir de sequências curadas de organismos modelo para outras negligenciadas, de organismos menos estudados. Mais uma vez, a ferramenta gerada desse processo, chamada U-MAGE, foi disponibilizada aos usuários em interface *web*. Pretendemos manter o conteúdo disponível sempre atualizado e agregando cada vez mais funcionalidades. Em conjunto, esperamos que as ferramentas bioinformáticas aqui descritas auxiliem no agrupamento e propagação do conhecimento. De fato, vários estudos em andamento no Laboratório de Biodados já se beneficiam dessas novas ferramentas e ao mesmo tempo em que adicionam curadoria manual do desempenho, induzem melhorias por intermédio de sugestões e necessidades.

9. Referências

1. Sonnhammer ELL, Koonin EV: **Orthology, paralogy and proposed classification for paralog subtypes.** *Trends in genetics : TIG* 2002, **18**:619–20.

2. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic acids research* 1997, **25**:3389–402.

3. Ostlund G, Schmitt T, Forslund K, Köstler T, Messina DN, Roopra S, Frings O, Sonnhammer ELL: **InParanoid 7: new algorithms and tools for eukaryotic orthology analysis.** *Nucleic acids research* 2010, **38**:D196–203.

4. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA: **The COG database: an updated version includes eukaryotes.** *BMC bioinformatics* 2003, **4**:41.

5. Waterhouse RM, Zdobnov EM, Tegenfeldt F, Li J, Kriventseva EV: **OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011.** *Nucleic acids research* 2011, **39**:D283–8.

6. Alexeyenko A, Tamas I, Liu G, Sonnhammer ELL: **Automatic clustering of orthologs and inparalogs shared by multiple proteomes.** *Bioinformatics (Oxford, England)* 2006, **22**:e9–15.

7. Roth ACJ, Gonnet GH, Dessimoz C: **Algorithm of OMA for large-scale orthology inference.** *BMC bioinformatics* 2008, **9**:518.

8. Li L, Stoeckert C, Roos D: **OrthoMCL: identification of ortholog groups for eukaryotic genomes**. *Genome research* 2003:2178–2189.

9. van der Heijden RTJM, Snel B, van Noort V, Huynen MA: **Orthology prediction at scalable resolution by phylogenetic tree analysis.** *BMC bioinformatics* 2007, **8**:83.

10. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E: **EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates.** *Genome research* 2009, **19**:327–35.

11. Datta RS, Meacham C, Samad B, Neyer C, Sjölander K: **Berkeley PHOG: PhyloFacts orthology group prediction web server.** *Nucleic acids research* 2009, **37**:W84–9.

12. Li H, Coghlan A, Ruan J, Coin LJ, Hériché J-K, Osmotherly L, Li R, Liu T, Zhang Z, Bolund L, Wong GK-S, Zheng W, Dehal P, Wang J, Durbin R: **TreeFam: a curated database of phylogenetic trees of animal gene families.** *Nucleic acids research* 2006, **34**:D572–80.

13. Altenhoff AM, Dessimoz C: **Phylogenetic and functional assessment of orthologs inference projects and methods.** *PLoS computational biology* 2009, **5**:e1000262.

14. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M: **The KEGG resource for deciphering the genome.** *Nucleic acids research* 2004, **32**:D277–80.

15. Barbosa-Silva A, Satagopam VP, Schneider R, Ortega JM: **Clustering of cognate proteins among distinct proteomes derived from multiple links to a single seed sequence.** *BMC bioinformatics* 2008, **9**:141.

16. Fernandes GR, Barbosa DVC, Prosdocimi F, Pena IA, Santana-Santos L, Coelho Junior O, Barbosa-Silva A, Velloso HM, Mudado MA, Natale DA, Faria-Campos AC, Aguiar SCV, Ortega JM: **A procedure to recruit members to enlarge protein family databases--the building of UECOG (UniRef-Enriched COG Database) as a model.** *Genetics and molecular research : GMR* 2008, **7**:910–24.

17. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M: **KEGG for integration and interpretation of large-scale molecular data sets.** *Nucleic acids research* 2012, **40**:D109–14.

18. Fernandes GR, Ortega JM: **Integração de bases de dados de genes homólogos e aplicações em análises de sequências.** 2011.

19. **The Universal Protein Resource (UniProt).** *Nucleic acids research* 2007, **35**:D193–7.

20. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH: **UniRef: comprehensive and non-redundant UniProt reference clusters.** *Bioinformatics (Oxford, England)* 2007, **23**:1282–8.

21. Mi H, Muruganujan A, Thomas PD: **PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees.** *Nucleic acids research* 2013, **41**:D377–86.

22. Maere S, Heymans K, Kuiper M: **BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks.** *Bioinformatics (Oxford, England)* 2005, **21**:3448–9.

23. McCarthy FM, Bridges SM, Wang N, Magee GB, Williams WP, Luthe DS, Burgess SC: **AgBase: a unified resource for functional analysis in agriculture.** *Nucleic acids research* 2007, **35**:D599–603.

24. Conesa A, Götz S: **Blast2GO: A comprehensive suite for functional analysis in plant genomics.** *International journal of plant genomics* 2008, **2008**:619832.

25. Mutowo-Meullenet P, Huntley RP, Dimmer EC, Alam-Faruque Y, Sawford T, Jesus Martin M, O'Donovan C, Apweiler R: **Use of Gene Ontology Annotation to understand the peroxisome proteome in humans.** *Database : the journal of biological databases and curation* 2013, **2013**:bas062.

26. Reeds PJ, Wahle KW, Haggarty P: **Energy costs of protein and fatty acid synthesis.** *The Proceedings of the Nutrition Society* 1982, **41**:155–9.

27. Aoyagi Y, Tasaki I, Okumura J, Muramatsu T: **Energy cost of whole-body protein synthesis measured in vivo in chicks.** *Comparative biochemistry and physiology. A, Comparative physiology* 1988, **91**:765–8.

28. Millward DJ: **Metabolic demands for amino acids and the human dietary requirement: Millward and rRvers (1988) revisited.** *The Journal of nutrition* 1998, **128**:2563S–2576S.

29. Millward DJ, Rivers JP: **The nutritional role of indispensable amino acids and the metabolic basis for their requirements.** *European journal of clinical nutrition* 1988, **42**:367–93.

30. Elango R, Ball RO, Pencharz PB: **Amino acid requirements in humans: with a special emphasis on the metabolic availability of amino acids.** *Amino acids* 2009, **37**:19–27.

31. Payne SH, Loomis WF: **Retention and loss of amino acid biosynthetic pathways based on analysis of whole-genome sequences.** *Eukaryotic cell* 2006, **5**:272–6.

32. Frishman D, Argos P: **Seventy-five percent accuracy in protein secondary structure prediction.** *Proteins* 1997, **27**:329–35.

33. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic acids research* 1994, **22**:4673–80.

34. Zemla A, Venclovas C, Fidelis K, Rost B: **A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment.** *Proteins* 1999, **34**:220–3.

35. Löytynoja A, Goldman N: **An algorithm for progressive multiple alignment of sequences with insertions.** *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**:10557–62.

36. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods.** *Molecular biology and evolution* 2011, **28**:2731–9.

37. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Molecular biology and evolution* 1987, **4**:406–25.

38. Mi H, Dong Q, Muruganujan A, Gaudet P, Lewis S, Thomas PD: **PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium.** *Nucleic acids research* 2010, **38**:D204–10.

39. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y: **KEGG for linking genomes to life and the environment.** *Nucleic acids research* 2008, **36**:D480–4.

40. Pollastri G, Przybylski D, Rost B, Baldi P: **Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles.** *Proteins* 2002, **47**:228–35.

41. Stathopoulos C, Jacquin-Becker C, Becker HD, Li T, Ambrogelly A, Longman R, Söll D: **Methanococcus jannaschii prolyl-cysteinyl-tRNA synthetase possesses overlapping amino acid binding sites.** *Biochemistry* 2001, **40**:46–52.

42. Nagel GM, Doolittle RF: **Evolution and relatedness in two aminoacyl-tRNA synthetase families.** *Proceedings of the National Academy of Sciences of the United States of America* 1991, **88**:8121–5.

43. Anselme J, Härtlein M: **Asparaginyl-tRNA synthetase from Escherichia coli has significant sequence homologies with yeast aspartyl-tRNA synthetase**. *Gene* 1989, **84**:481–485.

44. Berthet-Colominas C, Seignovert L, Härtlein M, Grotli M, Cusack S, Leberman R: **The crystal structure of asparaginyl-tRNA synthetase from Thermus thermophilus and its complexes with ATP and asparaginyl-adenylate: the mechanism of discrimination between asparagine and aspartic acid.** *The EMBO journal* 1998, **17**:2947–60.

45. Donnard E, Barbosa-Silva A: **Preimplantation development regulatory pathway construction through a text-mining approach**. *BMC …* 2011, **12 Suppl 4**:S3.

46. Katinka MD, Duprat S, Cornillot E, Méténier G, Thomarat F, Prensier G, Barbe V, Peyretaillade E, Brottier P, Wincker P, Delbac F, El Alaoui H, Peyret P, Saurin W, Gouy M, Weissenbach J, Vivarès CP: **Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi.** *Nature* 2001, **414**:450–3.

47. Miyazaki T, Miyazaki J, Yamane H, Nishiyama M: **alpha-Aminoadipate aminotransferase from an extremely thermophilic bacterium, Thermus thermophilus.** *Microbiology (Reading, England)* 2004, **150**:2327–34.

48. Velasco AM, Leguina JI, Lazcano A: **Molecular evolution of the lysine biosynthetic pathways.** *Journal of molecular evolution* 2002, **55**:445–59.

49. Hudson AO, Bless C, Macedo P, Chatterjee SP, Singh BK, Gilvarg C, Leustek T: **Biosynthesis of lysine in plants: evidence for a variant of the known bacterial pathways.** *Biochimica et biophysica acta* 2005, **1721**:27–36.

50. Nishida H, Nishiyama M, Kobashi N, Kosuge T, Hoshino T, Yamane H: **A prokaryotic gene cluster involved in synthesis of lysine through the amino adipate pathway: a key to the evolution of amino acid biosynthesis.** *Genome research* 1999, **9**:1175–83.

51. Smith TJ, Schmidt T, Fang J, Wu J, Siuzdak G, Stanley CA: **The structure of apo human glutamate dehydrogenase details subunit communication and allostery.** *Journal of molecular biology* 2002, **318**:765–77.

52. Allen A, Kwagh J, Fang J, Stanley CA, Smith TJ: **Evolution of glutamate dehydrogenase regulation of insulin homeostasis is an example of molecular exaptation.** *Biochemistry* 2004, **43**:14431–43.

53. Andersson JO, Roger AJ: **Evolution of glutamate dehydrogenase genes: evidence for lateral gene transfer within and between prokaryotes and eukaryotes.** *BMC evolutionary biology* 2003, **3**:14.

54. Altenhoff AM, Schneider A, Gonnet GH, Dessimoz C: **OMA 2011: orthology inference among 1000 complete genomes.** *Nucleic acids research* 2011, **39**:D289–94.

55. Geourjon C, Combet C, Blanchet C, Deléage G: **Identification of related proteins with weak sequence identity using secondary structure information.** *Protein science : a publication of the Protein Society* 2001, **10**:788–97.

56. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures.** *Journal of molecular biology* 1995, **247**:536–40.

57. Schaefer C, Schlessinger A, Rost B: **Protein secondary structure appears to be robust under in silico evolution while protein disorder appears not to be.** *Bioinformatics (Oxford, England)* 2010, **26**:625–31.

58. Dessimoz C, Gabaldón T: **Toward community standards in the quest for orthologs**. … 2012, **28**:900–904.

59. Nehrt NL, Clark WT, Radivojac P, Hahn MW: **Testing the ortholog conjecture with comparative functional genomic data from mammals.** *PLoS computational biology* 2011, **7**:e1002073.

60. Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C: **Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs.** *PLoS computational biology* 2012, **8**:e1002514.

61. Forslund K, Pekkari I, Sonnhammer ELL: **Domain architecture conservation in orthologs.** *BMC bioinformatics* 2011, **12**:326.

62. Peterson ME, Chen F, Saven JG, Roos DS, Babbitt PC, Sali A: **Evolutionary constraints on structural similarity in orthologs and paralogs.** *Protein science : a publication of the Protein Society* 2009, **18**:1306–15.

63. Movahedi S, Van de Peer Y, Vandepoele K: **Comparative network analysis reveals that tissue specificity and gene function are important factors influencing the mode of expression evolution in Arabidopsis and rice.** *Plant physiology* 2011, **156**:1316–30.

64. Guedes RLM, Prosdocimi F, Fernandes GR, Moura LK, Ribeiro H a L, Ortega JM: **Amino acids biosynthesis and nitrogen assimilation pathways: a great genomic deletion during eukaryotes evolution.** *BMC genomics* 2011, **12 Suppl 4**:S2.

65. Inglis DO, Skrzypek MS, Arnaud MB, Binkley J, Shah P, Wymore F, Sherlock G: **Improved Gene Ontology Annotation for Biofilm Formation, Filamentous Growth, and Phenotypic Switching in Candida albicans.** *Eukaryotic cell* 2013, **12**:101–8.

66. Pascovici D, Keighley T, Mirzaei M, Haynes PA, Cooke B: **PloGO: plotting gene ontology annotation and abundance in multi-condition proteomics experiments.** *Proteomics* 2012, **12**:406–10.

67. du Plessis L, Skunca N, Dessimoz C: **The what, where, how and why of gene ontology--a primer for bioinformaticians.** *Briefings in bioinformatics* 2011, **12**:723–35.

68. Skunca N, Altenhoff A, Dessimoz C: **Quality of computationally inferred gene ontology annotations.** *PLoS computational biology* 2012, **8**:e1002533.

10. Produção científica durante o Doutorado

10.1. Artigos científicos publicados em revistas internacionais

- **Amino acids biosynthesis and nitrogen assimilation pathways: a great genomic deletion during eukaryotes evolution.** **Guedes RL**, Prosdocimi F, Fernandes GR, Moura LK, Ribeiro HA, Ortega JM. BMC Genomics. 2011 Dec 22;12 Suppl 4:S2. PMID: 22369087.

- **Preimplantation development regulatory pathway construction through a text-mining approach**: Donnard E, Barbosa-Silva A, **Guedes RL**, Fernandes GR, Velloso H, Kohn MJ, Andrade-Navarro MA, Ortega JM. BMC Genomics. 2011 Dec 22;12 Suppl 4:S3. PMID: 22369103.

10.2. Capítulos de livros

ORTEGA, J. M. ; **GUEDES, R. L. M.** ; Abreu V. A. C. . Introdução ao Linux. In: Vasco Azevedo; Maria Paula Schneider; Artur Costa da Silva; Anderson Miyoshi; Aluízio Borém. (Org.). Manual Prático-Teórico – Sequenciamento, Montagem e Anotação de Genomas Bacterianos. 1ed.: , 2011, v. 1, p. 17-38

10.3. Trabalhos apresentados em congressos

COSTA, V. R. M. ; **Guedes, RLM** ; Ribeiro, HAL ; Ortega, JM . Depicting the origin of interferon, its receptors and regulatory transcription factors. 58º Congresso Brasileiro de Genética, 2012, Foz do Iguacu. *** Prêmio de melhor apresentação oral *** 2012.

VELLOSO., H. ; **GUEDES, R. L. M.** ; Ortega, JM . TaxSimple, LCAWS and GetAllChildren to make access to NCBI Taxonomy data easier and faster. X-meeting 2012. 8th International Conference of the Brazilian Association for Bioinformatics and Computational Biology, 2012, Campinas, 2012.

COSTA, V. R. M. ; **GUEDES, R. L. M.** ; ORTEGA, J. M. . The rise of cell-to-cell contacts in the presentation of antigens to T-cells in human evolution. X-meeting 2012. 8th International Conference of the Brazilian Association for Bioinformatics and Computational Biology, 2012, Campinas. *** Prêmio de menção honrosa *** 2012.

Stussi F. ; **GUEDES, R. L. M.** ; Barbosa-Silva, Adriano ; ORTEGA, J. M. . Depicting the origin of sumoylation network in Arabdopsis thaliana with text-mining and orthologue clustering tools. X-meeting 2012. 8th International Conference of the Brazilian Association for Bioinformatics and Computational Biology, 2012, Campinas, 2012.

GUEDES, R. L. M. ; NATALE, D. ; SUZEK, B. E. ; WU, C. H. ; ORTEGA, J. M. . UniRef50 matrices for Annotating Gene Ontology Entries (U-MAGE) web tool. 1st ICBI Biomedical Informatics Symposium at Georgetown University 2012, Washington DC. 2012.

SAKAMOTO, T. ; COSTA, V. R. M. ; KOLMAN, M. ; **GUEDES, R. L. M.** ; GIANI, A. ; Ortega, JM . Occurence and phylogenetic studies of the microcystin biosynthesis genes. X-meeting 2012. 8th International Conference of the Brazilian Association for Bioinformatics and Computational Biology, 2012, Campinas, 2012.

FERREIRA, L. ; **GUEDES, R. L. M.** ; ORTEGA, J. M. . How old are the genes in E. coli operons? X-meeting 2012. 8th International Conference of the Brazilian Association for Bioinformatics and Computational Biology, 2012, Campinas, 2012.

GUEDES, R. L. M. ; NATALE, D. ; SUZEK, B. E. ; WU, C. H. ; Ortega, JM . UniRef50 matrices for Annotating Gene Ontology Entries (U-MAGE) among species in PANTHER database. X-meeting 2012. 8th International Conference of the Brazilian Association for Bioinformatics and Computational Biology, 2012, Campinas, 2012.

GUEDES, R. L. M. ; Coelho-Júnior, O. ; VELLOSO., H. ; RIBEIRO, H. A. L. ; FERNANDES G R ; DONNARD, E. ; Stussi F. ; ORTEGA, J. M. . The secondary structure of orthogues is nearly 90% versus 70% similar if they respectively belong to the same species or phyla. X-meeting 2011. 7th International Conference of the Brazilian Association for Bioinformatics and Computational Biology, 2011, Florianopolis, 2011.

Stussi F. ; **GUEDES, R. L. M.** ; VELLOSO., H. ; DONNARD, E. ; RIBEIRO, H. A. L. ; FERNANDES G R ; Barbosa-Silva, A. ; Andrade-Navarro, M. A. ; ORTEGA, J. M. . Depicting the origin of drought response in Arabdopsis thaliana with text-mining and orthologue clustering tools. X-meeting 2011. 7th International Conference of the Brazilian Association for Bioinformatics and Computational Biology, 2011, Florianopolis. *** Prêmio de melhor poster ***, 2011.

VELLOSO., H. ; URBINA H. ; RIBEIRO, H. A. L. ; HONORATO R. ; **GUEDES, R. L. M.** ; PEREZ-ACLE T. ; ORTEGA, J. M. . BOWS: a platform to make bioinformatics tools available by Web Service. X-meeting 2011. 7th International Conference of the Brazilian Association for Bioinformatics and Computational Biology, 2011, Florianópolis, 2011.

ALVARENGA E. R. ; MAGALHAES B. F ; DANTAS A. E. ; SIQUEIRA F. F. ; MENDES T.M. ; **GUEDES, R. L. M.** ; ORTEGA, J. M. ; KALAPOTHAKIS, E. . Comparative trascriptomic of Chelicerates: searching for gene expression pattern in venom glands. X-meeting 2011. 7th International Conference of the Brazilian Association for Bioinformatics and Computational Biology, 2011, Florianópolis, 2011.

GUEDES, R. L. M. ; VELLOSO., H. ; ORTEGA, J. M. . Metagenomic Clustering for Environmental Analyses. SB-Meeting, Brazil Deutschland Systems Biology Meeting, 2010, Ouro Preto. Metagenomic Clustering for Environmental Analyses, 2010.

DONNARD, E. ; **GUEDES, R. L. M.** ; VELLOSO., H. ; FERNANDES G R ; Coelho-Júnior, O. ; RIBEIRO, H. A. L. ; Barbosa-Silva, A. ; ORTEGA, J. M. . Preimplantation development Regulatory Pathway : searching for homologues and their last common ancestor. SB-Meeting, Brazil Deutschland Systems Bilogy Meeting, 2010, Ouro Preto. Preimplantation development Regulatory Pathway : searching for homologues and their last common ancestor, 2010.

VELLOSO., H. ; **GUEDES, R. L. M.** ; ORTEGA, J. M. . Genes Seems to Have Appeared in Waves Algong the Evolution. SB-Meeting, Brazil Deutschland Systems Bilogy Meeting, 2010, Ouro Preto. Genes Seems to Have Appeared in Waves Algong the Evolution, 2010.

GUEDES, R. L. M. ; Moura, L.K. ; Queiroz, L.S. ; VELLOSO., H. ; ORTEGA, J. M. ; PROSDOCIMI, F. . The Great Deletion Hypothesis for Essential Amino Acid Biosynthesis Enzymes. XXXIX Reunião Anual da Sociedade Brasileira de Bioquímica e Biologia Molecular - SBBq, 2010, Foz do Iguaçu. The Great Deletion Hypothesis for Essential Amino Acid Biosynthesis Enzymes, 2010.

DONNARD, E. ; **GUEDES, R. L. M.** ; VELLOSO., H. ; fernandes G R ; Coelho-Júnior, O. ; RIBEIRO, H. A. L. ; Barbosa-Silva, A. ; ORTEGA, J. M. . Preimplantation Development Consists of an Ancient Chordata Transcriptional Regulatory Pathway With Addition of Modern Elements by the Origin of Eutheria. XXXIX Reunião Anual da Sociedade Brasileira de Bioquímica e Biologia Molecular - SBBq, 2010, Foz do Iguaçu.

Stussi F. ; Barbosa-Silva, A. ; DONNARD, E. ; fernandes G R ; **GUEDES, R. L. M.** ; ORTEGA, J. M. . Enriching KEGG Pathway With PESCADOR Text Mining Tool - Colorectal Cancer Pathway As a Model. X-meeting 2010. 6th International Conference of the Brazilian Association for Bioinformatics and Computational Biology, 2010, Ouro Preto. X-meeting 2010. 6th International Conference of the Brazilian Association for Bioinformatics and Computational Biology, 2010.

RIBEIRO, C. ; **GUEDES, R. L. M.** ; RIBEIRO, H. A. L. ; Coelho-Júnior, O. ; ORTEGA, J. M. . Design of a Quintillion of Genes that Codify for Synthetic mini-antibodies (ScFv). X-Meeting 2010 - AB C 6th International Conference of the Brazilian Association for Bioinformatics and Computational Biology., 2010, Ouro Preto. X-Meeting 2010 - AB C 6th International Conference of the Brazilian Association for Bioinformatics and Computational Biology, 2010.

GUEDES, R. L. M. ; VELLOSO, H. M. ; ORTEGA, J. M. . One quarter of all metagenomics sequences can be clustered but only a small fraction can have a genus assigned due to the limitation of sequenced genomes. 55º Congresso Brasileiro de Genética, 2009, Aguas de Lindoia - SP. One quarter of all metagenomics sequences can be clustered but only a small fraction can have a genus assigned due to the limitation of sequenced genomes, 2009.

11.Anexo I

Guedes et al. BMC Genomics 2011, 12(Suppl 4):S2
http://www.biomedcentral.com/1471-2164/12/S4/S2

BMC Genomics

PROCEEDINGS Open Access

Amino acids biosynthesis and nitrogen assimilation pathways: a great genomic deletion during eukaryotes evolution

RLM Guedes[1], F Prosdocimi[2,3], GR Fernandes[1,2], LK Moura[2], HAL Ribeiro[1], JM Ortega[1*]

From 6th International Conference of the Brazilian Association for Bioinformatics and Computational Biology (X-meeting 2010)
Ouro Preto, Brazil. 15-18 November 2010

Abstract

Background: Besides being building blocks for proteins, amino acids are also key metabolic intermediates in living cells. Surprisingly a variety of organisms are incapable of synthesizing some of them, thus named Essential Amino Acids (EAAs). How certain ancestral organisms successfully competed for survival after losing key genes involved in amino acids anabolism remains an open question. Comparative genomics searches on current protein databases including sequences from both complete and incomplete genomes among diverse taxonomic groups help us to understand amino acids auxotrophy distribution.

Results: Here, we applied a methodology based on clustering of homologous genes to seed sequences from autotrophic organisms *Saccharomyces cerevisiae* (yeast) and *Arabidopsis thaliana* (plant). Thus we depict evidences of presence/absence of EAA biosynthetic and nitrogen assimilation enzymes at phyla level. Results show broad loss of the phenotype of EAAs biosynthesis in several groups of eukaryotes, followed by multiple secondary gene losses. A subsequent inability for nitrogen assimilation is observed in derived metazoans.

Conclusions: A Great Deletion model is proposed here as a broad phenomenon generating the phenotype of amino acids essentiality followed, in metazoans, by organic nitrogen dependency. This phenomenon is probably associated to a relaxed selective pressure conferred by heterotrophy and, taking advantage of available homologous clustering tools, a complete and updated picture of it is provided.

Background

Creation and analysis of groups of orthologous genes have been widely used for gene function prediction, evolutionary and divergence time studies [1]. Moreover, orthology is also a valuable source for evolutionary comprehension of pathways through phylogenetic analysis. In respect to a central issue on cellular metabolism, the order of appearance for universal cellular metabolisms was estimated by Cunchillos and Lecointre [2,3], with amino acid catabolism and anabolism being respectively the first and second pathways to appear, even earlier than glycolysis and gluconeogenesis. The amino acids biosynthesis, rather than linear and universal series of reactions with homologues occurring in different organisms, sometimes relies on alternative pathways, as shown by Hernández-Montes et al. [4]. Moreover, gene loss and pathway depletion, important events in genome evolution, can be inferred from the orthologous groups through comparative genomics. Today, a vast amount of information is provided by intensive genome sequencing, and the efforts of grouping homologous genes had reached great standards.

Amino acid anabolism is responsible for about 20% of the energy that cells spend on protein synthesis [5,6]. The nutritional requirements of essential amino acids and nitrogen are of striking importance and they have

* Correspondence: miguel@icb.ufmg.br
[1]Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, 31270-901, MG, Brazil
Full list of author information is available at the end of the article

been estimated as ~22mg/kg of EAAs and 3mg/kg of N in human body [7,8]. More recent approaches for dietary requirement calculations, using amino acid oxidation as an indicator, reveal that the requirement is over five fold what the classical approaches indicated, and the requirement has now been determined for each of the nine human EAAs [9]. It is of general understanding that plant, as well as fungi, synthesize all amino acids required for protein synthesis and that evolutionary processes culminated in human inability to synthesize nine amino acids (histidine, phenylalanine, tryptophan, valine, isoleucine, leucine, lysine, methionine and threonine), thus called essential amino acids (EAAs), which must be obtained through diet. Amino acids also constitute our source of organic nitrogen. There have been few attempts to understand why some amino acids have become essential. However, genome deletion events have happened in the past and many organisms have lost a number of important enzymes necessary for *de novo* biosynthetic pathways. Hitherto, the pattern of loss versus retention for amino acids biosynthetic pathways was analyzed for a few protists and metazoans by Payne and Loomis [10]. They verified that the set of essential amino acids is the same in animals and protists. Curiously, most of the retained amino acids are intermediates in secondary pathways like purine ring biosynthesis and nitrogen metabolism.

An overview for the presence/absence of the enzymes which compose the amino acid biosynthetic pathways, among distinct phyla in the tree of life, could be accomplished with (i) rich protein databases such as the UniProt Knowledgebase (UniProtKB) [11] comprising over 10 million full-length sequences and (ii) the current initiatives to group these proteins by evolutionary relatedness - called homologues - such as COG-Cluster of Orthologous Groups [12] and KEGG Orthology [13]. Unfortunately these initiatives consider only proteins derived from complete genomes and thus a large amount of information is currently lost, with over 6 million remaining full-length proteins that belong to organisms with still incomplete genomes.

Here, we applied a methodology that takes into account all available protein information to depict, at phyla level, the EAA biosynthetic and nitrogen assimilation enzymes scenarios to inspect how and when amino acid auxotrophy has first appeared along evolution.

A Great Genomic Deletion model is proposed to explain the phenotypic inability to synthesize amino acids that appears independently in distinct phylogenetically distant clades of eukaryotes. Such events should be followed by subsequent steps of gene loss due to relaxed selective pressure in already incomplete pathways, leading to an eventual loss of all genes for a particular biosynthesis pathway in some clades. Accordingly, in metazoans but Cnidaria, dependence on organic nitrogen accompanies the evolution of heterotrophy, thus organisms become dependent even on NEAA for supplying their nitrogen requirements.

Results

Clustering homologues of amino acid biosynthetic enzymes

To determine the distribution of amino acid biosynthetic enzymes, a homologue clustering process was developed to allow the use of both complete and incomplete genomes [14,15]. The procedure starts with Seed Linkage software [14] that clusters cognate proteins from multiple organisms beginning with a single seed sequence through connectivity saturation with it. Since basal eukaryotes such as plants and fungi are autotrophic, sequences coding for all the enzymes used in the biosynthesis of EAAs from the plant *Arabidopsis thaliana* and the fungus *Saccharomyces cerevisiae* were manually inspected using KEGG Pathway and used as seeds to search for homologues. Moreover, our group has been developing a procedure to enrich secondary databases such as COG [12] and KEGG Orthology (to be published) with UniRef50 clusters [16] available from UniProt, therefore allowing the inclusion of data from incompletely sequenced genomes. Additional file 1: Sequences and genome status distribution reflects the abundance of proteins derived from incomplete genomes and evidences the importance of their inclusion. In this work we took advantage of a home-built UniRef50 Enriched KEGG Orthology database (UEKO) to additionally cluster sequences with the seed sequences mentioned above. Since these searches recruit sequences from diverse clades, which may or may not contain organisms with completely sequenced genomes, we represented this information in Figure 1 as: (a) black filled circles for phyla containing complete genomes; (b) grey filled circles comprise clades with at least one draft genome available, but no complete genome, and (c) empty circles represent phyla with no complete nor draft genomes. Protein fragments are not included in the search for homologues because they may represent partial sequenced full length proteins at mRNA level or incompletely modeled from genome. Moreover since some full length proteins might have not been captured in databases due to high sequence divergence, a second search round used UniProt to query all clustered sequences. This step also captures partial sequences (entries labeled as fragments in UniProt) which were approved by the coverage filtering applied (see Methods for details). These additional significant hits are represented by triangles in Figure 1. Furthermore, enzymes required for the biosynthesis of the indicated amino acids are ordered in the anabolic pathway from left to

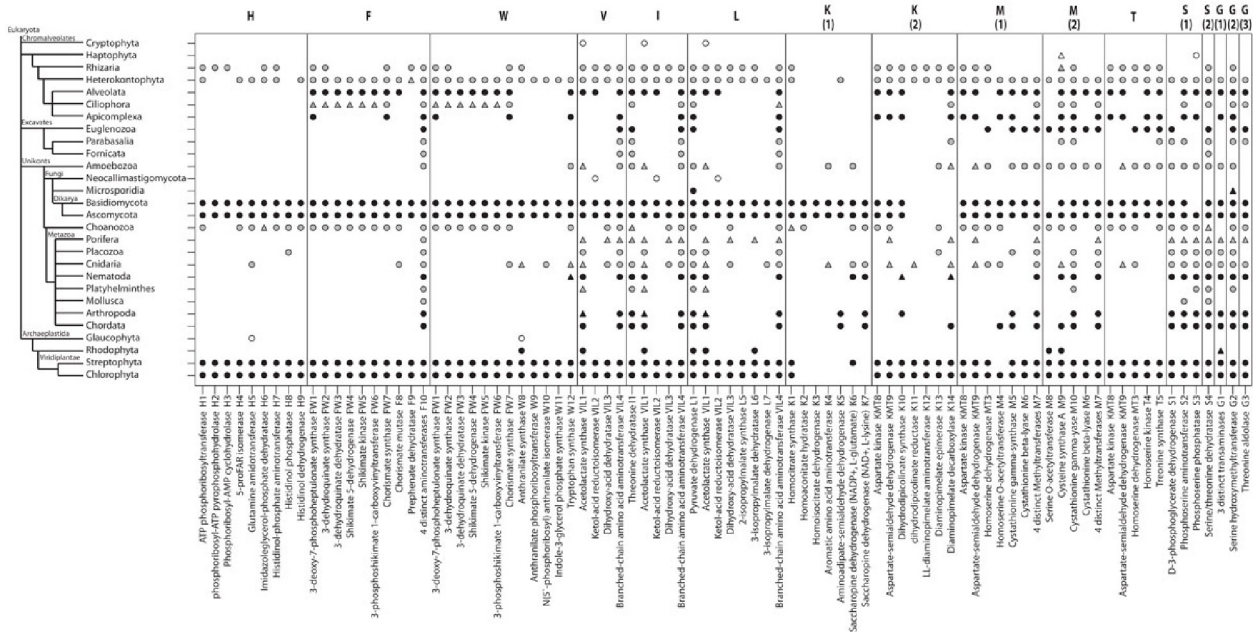


Figure 1 Essential amino acid anabolic pathways. Schematic representation for presence/absence of anabolic enzymes for nine essential amino acids and the non-essential amino acids serine and glycine. Eukaryotic taxonomic tree displayed at phyla level. Circles represent detection of complete proteins and triangles detection of complete and fragmented proteins. Black: phyla containing complete genomes; Grey: at most organisms with draft genomes; White: phyla with no complete or draft genomes. *Saccharomyces cerevisiae* (Ascomycota) and *Arabidopsis thaliana* (Streptophyta) were used as seeds. The 4 distinct aminotransferases in phenylalanine pathway are: (i) aspartate aminotransferase (ii) histidinol-phosphate aminotransferase (iii) aromatic amino acid aminotransferase (iv) tyrosine aminotransferase. The 4 distinct methyltransferases in methionine pathway are: (i) 5-methyltetrahydropteroyltriglutamate–homocysteine methyltransferase (ii) homocysteine S-methyltransferase (iii) betaine-homocysteine methyltransferase (iv) 5-methyltetrahydrofolate–homocysteine methyltransferase. The 3 distinct transaminases in glycine pathway are: alanine-glyoxylate transaminase, serine-glyoxylate transaminase and serine-pyruvate transaminase.

right. All pathways refer to EAAs biosynthesis except serine and glycine (the rightmost ones) used as experimental controls. Serine is represented with two alternative pathways observed in human and other eukaryotes: S(1), from 3P-D-glycerate; and S(2), from pyruvate. Glycine is also represented by two pathways: G(1) and G(2), both coming from serine; and G(3), coming from threonine. As expected, serine and glycine biosynthesis were found to be potentially proficient in almost all phyla. This control supports the searching mechanism and attest for the efficacy of methods applied. A few exceptions were observed and deserve comments: (i) Serine biosynthetic pathways was found to be absent in Rhodophyta, although the complete genome of *Cyanidioschyzon merolae* is available. We manually inspected this result with regular BLAST searches and did not find additional evidence, although a translation of partial CDS was obtained for glycine biosynthetic enzyme G1 (Figure 1, triangle); (ii) Serine biosynthesis seems absent in Apicomplexa as well, a clade comprising two *Plasmodium* complete genomes lacking enzymes S1 and S4; (iii) Considering the animals, besides being able to find serine biosynthetic enzymes, we fail to support the NEAA character of glycine for Mollusca. However, evidences could be obtained for ancient organisms such as Placozoa and Porifera. For the Microsporidia *E. cuniculi*, an obligatory intracellular parasitic fungus with complete genome, it has been reported that "the repertoire for the biosynthesis of amino acids is restricted to asparagines synthetase and serine hydroxymethyltransferase genes", then serine was known as an EAA [17]. Thus, absence of evidence may not guarantee the absence of the gene. However, out of 28 phyla, discarding both the four clades with no genome project or in progress (open circles) and the ones with complete genome (filled symbols), we could not provide evidence of glycine biosynthesis for two phyla (Fornicata and Mollusca). However evidence for serine has been provided in all of them.

Data presented in Figure 1 clearly depicts the presence of complete biosynthetic pathways for EAAs in both plants (Chlorophyta and Streptophyta) and fungi (Ascomycota and Basidiomycota), as stated above. In previous work we hypothesized that a great event of genome deletion on which many of the intermediate enzymes for biosynthetic pathways for amino acids have vanished, ended up affecting the usage of EAAs in chordate proteomes [18,19]. In 2006, Payne and Loomis [10] using

pFam protein signatures reported that protists and animals share essentiality for the nine amino acids. Here we provide a broader analysis covering all genomes available today and trying to map how and when the Great Genomic Deletion has happened. Evidence was found suggesting that this loss of capability to synthesize EAAs is conspicuous at the base of metazoan evolution, simultaneously affecting the complete set of EAAs. The phenomenon is characterized as an initial phenotypic deficiency, observed in Choanozoa, followed by multiple secondary gene losses. Accordingly, some enzymes found in Chordata such as K14, M4 and M9 are missing in Arthropoda. Remarkably, some components such as VIL1 and M7 are maintained in most metazoan clades, despite of pathway loss.

Actually, a Great Deletion causing concurrent phenotypic loss of amino acid biosynthesis capability affects both metazoan and non-metazoan eukaryotes. Several clades containing complete genomes (black filled symbols) such as Rhodophyta, Euglenozoa and Apicomplexa, show similar EAAs pattern. Moreover, some evidence is provided suggesting the absence of complete pathways in the non-Dikarya Fungi Microsporidia and Neocallimastigomycota. This gives support to separate events of Great Genomic Deletion for the origin of EAAs auxotrophy in at least three other branches. Similarly to Choanozoa, clades such as Heterokontophyta and Rhizaria present various enzymes and some complete pathways. Evidences of complete pathways for all EAAs but histidine (H) were obtained in Heterokontophyta. Valine (V), isoleucine (I), lysine (K) and threonine (T) are potentially synthesized in Rhizaria as well as methionine (M) in Euglenozoa and Amoebozoa. However it is possible that other EAAs may also be synthesized in some of these clades. The anabolic capabilities suggested by the current data might be underestimated because we have only draft genomes available for most of these organisms. The Choanozoa clade contains only draft genomes. Though we observed more enzymes than in metazoan clades, a final picture of Choanozoan phenylalanine biosynthesis, for example, might require completion of genome sequencing. Further gene loss occurs during metazoan evolution; however, for Placozoa, Porifera and Cnidaria, the Great Genomic Deletion seems to be well established. Since the first available sponge genome is still an ongoing project and its proteins are not yet deposited in UniProt, we manually inspected the deduced proteome using regular BLAST alignments (see Methods) and evidenced auxotrophy for all nine EAAs. The same simple approach was applied to all phyla (Figure 1, triangles). Other clades that do not present any enzymes were omitted from Figure 1, such as Apusozoa and Jakobida.

Lysine biosynthesis

Inspection of Figure 1 depicts a remarkable difference on lysine (K) biosynthesis pathways present in fungi and plants. Since the occurrence of an α-aminoadipate (AAA) pathway K(1) in Fungi [20] as opposite to a diaminopimelate (DAP) pathway K(2) known to be present in plants, algae and bacteria [21,22] has already been reported, we set up to depict the complete scenario for K biosynthesis including prokaryotes (Figure 2). A third pathway K(3) preferentially used by Archaea but also reported to exist in bacterial groups [23] was also considered, therefore sequences from the *Pyrococcus horikoshii* archaea were also used as seed for homologue sequence clustering. Data supports the view that the K (2) pathway, found to be complete in plants, is often present in prokaryotic clades of bacteria and archaea, in agreement with previous findings [21,22]. Curiously, nine bacterial clades (Acidobacteria, Chlorobi, Deferribacteres, Deinococcus-Thermus, Fusobacteria, Chlamydiae, Synergistetes, Tenericutes and Thermotogae) – all of which contain complete genomes – do not present K12 enzyme, but there are three other alternative subsets of enzymes present in prokaryotes that could circumvent this step in lysine biosynthesis. Chlamydiae may represent an evidence of amino acid essentiality extended to prokaryotes, since diaminopimelate decarboxylase (K14) is absent and there are no known alternatives to this reaction. The set of enzymes responsible for the K(3) pathway, was found to occur in prokaryotes, and it is complete in the archaeal clades Crenarchaeota and Euryarcheota, as well as in the bacterial clades Chloroflexi and Proteobacteria, and probably in Actinobacteria and Bacteroidetes. Remarkably, the first four enzymes that constitute this pathway are coincident with the K(1) pathway (indicated by gray shading). The complete K(1) pathway occurs in Proteobacteria (and possibly in Actinobacteria, Bacteroidetes and Firmicutes, as evidenced by regular BLAST) and fungi. Thus, it is tempting to assume that a variant synthesis of K occurred in Archaea and, being modified in one of the four bacterial phyla above (with the addition of three enzymes: aminoadipate-semialdehyde dehydrogenase, saccharopine dehydrogenase NADP+ and saccharopine dehydrogenase NAD+), ended up constituting the fungi-occurring K biosynthetic pathway. The eukaryotic clades Rhizaria and Heterokontophyta, which present the K(2) pathway, appear to group with plants.

Nitrogen auxotrophy

Consumption of amino acids is an important route for nitrogen assimilation in other biological compounds for heterotrophic organisms, such as those comprised by some of the clades shown in Figure 1 (e.g. Chordata). Assimilation of free ammonium in eukaryotes is done

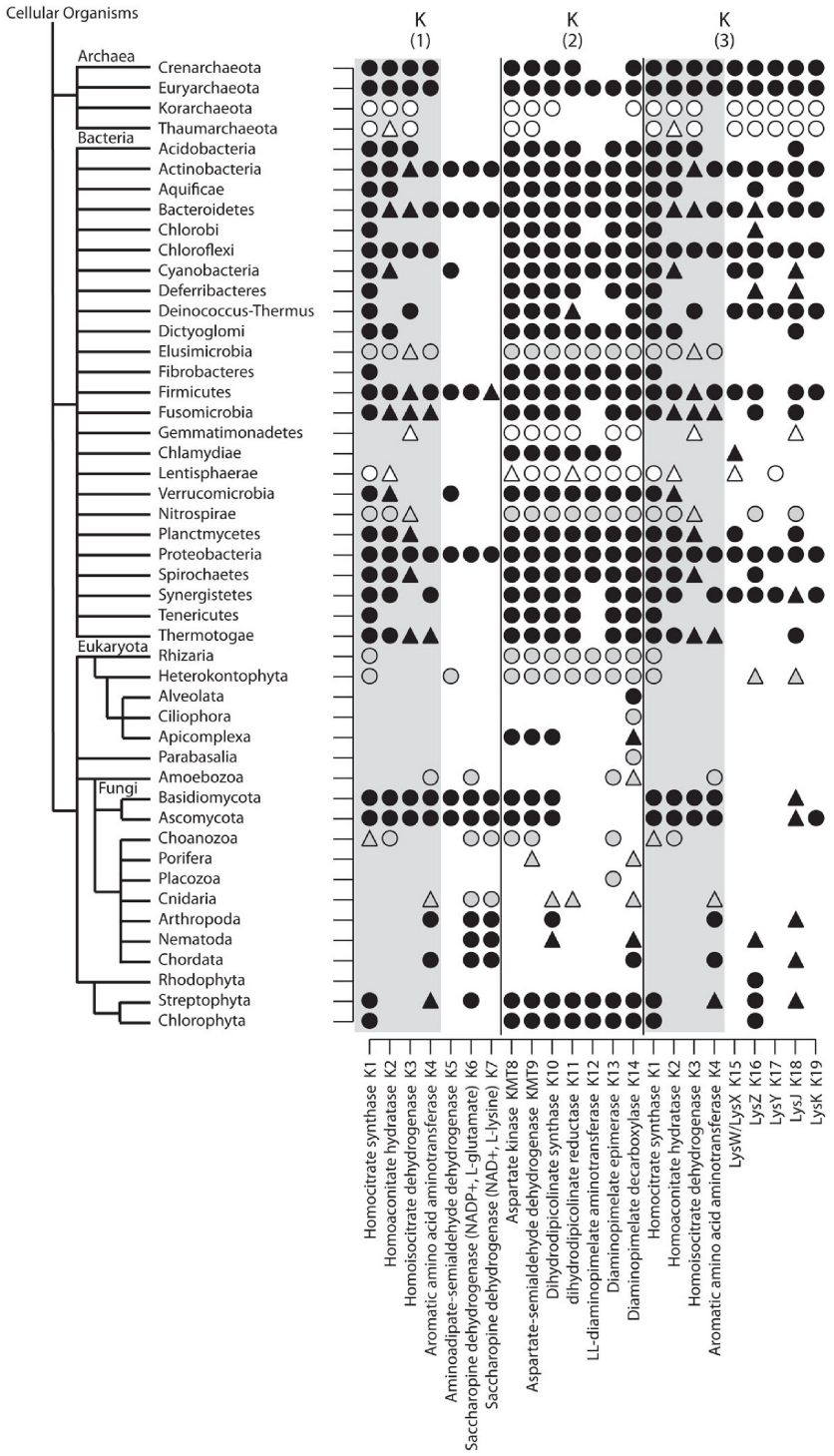


Figure 2 Lysine anabolic pathways. Schematic representation for presence/absence of enzymes involved in lysine biosynthesis. K(1) represents Fungi α-aminoadipate (AAA) pathway; K(2) bacteria, plants, and algae diaminopimelate (DAP) pathway; K(3) archaea α-aminoadipate (AAA) variant pathway. Taxonomic tree displayed at phyla level. Circles represent detection of complete proteins and triangles detection of complete and fragmented proteins. Colors are as for Figure 1. *Saccharomyces cerevisiae* (Ascomycota), *Arabidopsis thaliana* (Streptophyta) and *Pyrococcus horikoshii* (Euryarchaeota) were used as seeds.

by a cytoplasmatic reaction catalyzed by glutamate dehydrogenase (EC:1.4.1.4) which incorporates ammonium into alpha-ketoglutarate yielding glutamate, using electrons from a reduced cytoplasmatic co-enzyme NADPH. Two isoforms are present in fungi and one in plants, the latter having the additional option to not only assimilate nitrogen, but also to fixate it, often with the association of nitrogen-fixating bacteria. Thus, to investigate if the Great Genomic Deletion of biosynthetic enzymes for EAAs co-occurred with the heterotrophy for nitrogen, we generated clusters of the assimilative isoforms (EC:1.4.1.4) and, as a control, the mitochondrial enzymes (EC:1.4.1.2) which tend to operate in the reverse direction, i.e. glutamate degradation, by oxidizing it and delivering ammonium, loading electrons in NAD+ co-enzyme. In yeast, the cytoplasmic assimilative isoforms are named *GDH1* and *GDH3*, and the catabolic (mitochondrial) is known as *GDH2*. *Arabidopsis thaliana* proteins were also used as seed together with the *Saccharomyces cerevisiae* sequences: one known as putative *GDH* which grouped with the fungi assimilative ones, and three catabolic *GDHs*, that grouped with the human mitochondrial *GLUD1*, though not with the yeast catabolic *GHD2*. Results are shown in Figure 3A. The left column shows a cluster that groups assimilative isoforms with the two from yeast and the putative *GDH* from *A. thaliana*. The catabolic mitochondrial isoforms from yeast (central column) and plant (right column) formed two independent clusters. In metazoan organisms, an assimilative enzyme was found in the basal group Cnidaria, all others being dependent on amino acid consumption to build nitrogenated compounds such as DNA, Porifera included. Assimilative isoforms were also lacking in Choanozoa although complete genomes are unavailable. The same was observed for Placozoa. Comparing these results with those shown in Figure 1, it is remarkable that Choanozoa, while still registering many amino acid biosynthetic enzymes (37 out of 61, redundancy eliminated) shows a simultaneous deletion in both EAAs biosynthesis and nitrogen assimilation. It is also apparent that the Great Genomic Deletion attains its almost final broad distribution in Cnidaria, which may be the last metazoan clade still capable to assimilate nitrogen from free ammonium. Therefore a few biosynthetic enzymes remain, in this clade and other Metazoa, probably by connective functions in metabolism (e.g. EC: 1.2.1.31 aminoadipate-semialdehyde dehydrogenase K5 and EC: 1.5.1.7 saccharopine dehydrogenase K7 also participates in the lysine degradation pathway). We have also observed that mammalian *GDH* (*GLUD1*) presents a specialized allosteric control [24] which might have turned the enzyme toward glutamate catabolism rather than anabolism. Such control was first observed in Ciliophora [25] and it is thought to have been transferred by lateral gene transfer to the metazoan ancestor [26]. To confirm the grouping in three clusters of enzymes with so similar activities, Figure 3B shows a phylogenetic tree built with eukaryotic glutamate dehydrogenase sequences, which clustered the isoforms in total accordance with data shown in Figure 3A.

The non-Metazoa eukaryotes with complete genomes, such as Alveolata, Apicomplexa and Euglenozoa, lack EAA biosynthetic enzymes (Figure 1) but keep the capability of nitrogen assimilation (Figure 3). Fornicata and Parabasalia, although represented only by draft genomes, have shown to contain the nitrogen assimilation enzyme even if they appear to be auxotrophic for all EAAs. Lacking detection of any isoform of glutamate dehydrogenase and with available draft genomes is Rhizaria (no complete genomes available), which still presents some EAA biosynthetic capability. It is possible that the dependency of organic nitrogen has been attained earlier in Rhizaria, although complete sequencing is required for a sound conclusion. In general, data support a tendency for nitrogen heterotrophy succeeding the amino acid essentiality. In Rhodophyta, a clade containing complete genomes sequenced, surprisingly no catabolic homologues were found; however a sequence that clusters with the assimilative isoforms has been found.

We also investigated nitrogen assimilation in prokaryotes. Homologues of assimilative enzymes are present and detected by our clustering procedure, but besides finding homologues of the catabolic seeds in bacterial clades, assimilative enzymes were not found in Aquificae, Chlamydiae and Synergistetes, all of them containing complete genomes available. This absence is consistent with the lysine auxotrophy suggested in Chlamydiae (Figure 2) and support the idea that EAA auxotrophy is associated with the lack of nitrogen assimilation even in the prokaryotic clades. It is hard to infer differential enzymatic activity in prokaryotes, since the annotated sequences available often report mixed use of coenzyme, either NADPH or NAD, although the homologous tools had grouped them distinctively. If the homology is related to function, it may indicate that these organisms also demand the consumption of NEAA to constitute a source of organic nitrogen. The presented scenario suggests that the loss of nitrogen assimilation forcing consumption of NEAA shortly succeeds the Great Genomic Deletion of EAA biosynthetic enzymes in metazoans. If this hypothesis is true, the Cnidaria would be an exception.

EAA biosynthetic enzymes maintained

The remaining EAA biosynthetic enzymes in organisms that do not have the complete amino acid pathway (Figure 1) are more susceptible to evolutionary

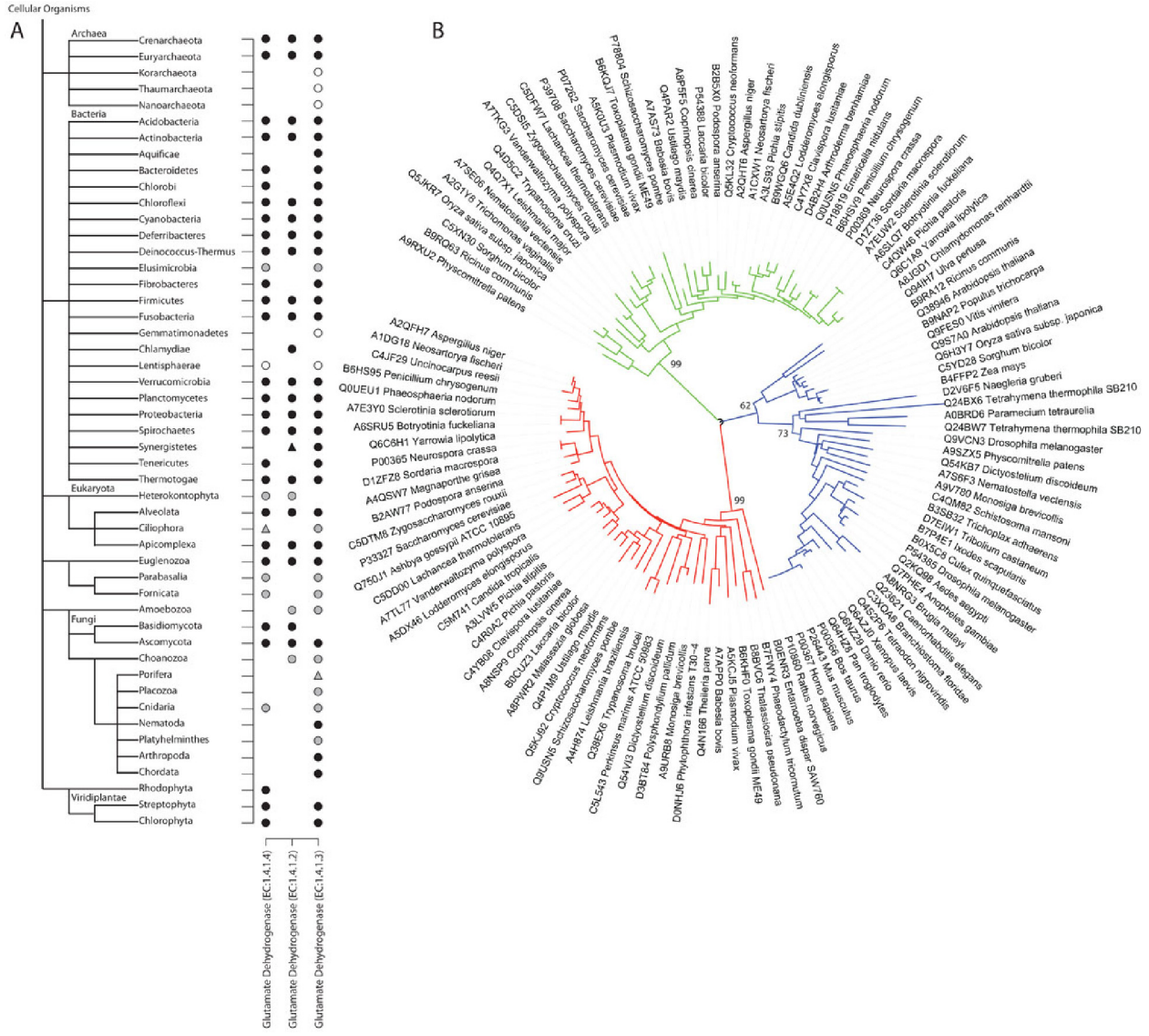


Figure 3 Glutamate dehydrogenases. Schematic representation for presence/absence of glutamate dehydrogenases. A: Left column: assimilative *GDH1* and *GDH3* from *Saccharomyces cerevisiae* and putative *GDH* from *Arabdopsis thaliana*; Central column: catabolic *GDH2* from *Saccharomyces cerevisiae*; Right column: catabolic *GDH1*, *GDH2* and *GDH3* from *Arabdopsis thaliana*. Taxonomic tree displayed at phyla level. Circles represent detection of complete proteins and triangles detection of complete and fragmented proteins. Colors are as for Figure 1. *Saccharomyces cerevisiae* (Ascomycota) and *Arabidopsis thaliana* (Streptophyta) were used as seeds. B: Phylogenetic tree with eukaryotic sequences from glutamate dehydrogenase isoforms. Green branches: EC1.4.1.4; Red branches: EC:1.4.1.2; Blue branches: EC:1.4.1.3.

modifications. It is also possible that paralogue subfunctionalization occurred in the common ancestor of animals, fungi and plants, and thus the divergent copy has remained in detriment of the original gene. Considering both hypothesis we set up to analyze enzymes from EAA and functional NEAA pathways present in metazoans. Phylogenetic trees for acetolactate synthase (VIL1 code in Figure 1) and for a group of alanine-glyoxylate, serine-glyoxylate and serine-pyruvate transaminases (G1 code in Figure 1) are represented in Figure 4. As expected, the distance between the ancestors of the two prototrophic groups varies, plant (green circles) and fungi (yellow circles): 0.4 and 0.7, for VIL1 (Figure 4A) and G1 (Figure 4B), respectively. The distance from the ancestors of plant (green circles) to metazoans (red circles) are relatively higher for the remaining enzyme VIL1: 1.0 (as compared to 0.4 measured from plant to fungi, 2.5 fold) than for the NEAA biosynthetic enzyme G1: 0.7 (as compared to 0.7 measured from plant to fungi, 1.0 fold). Thus, the remaining EAA enzymes are

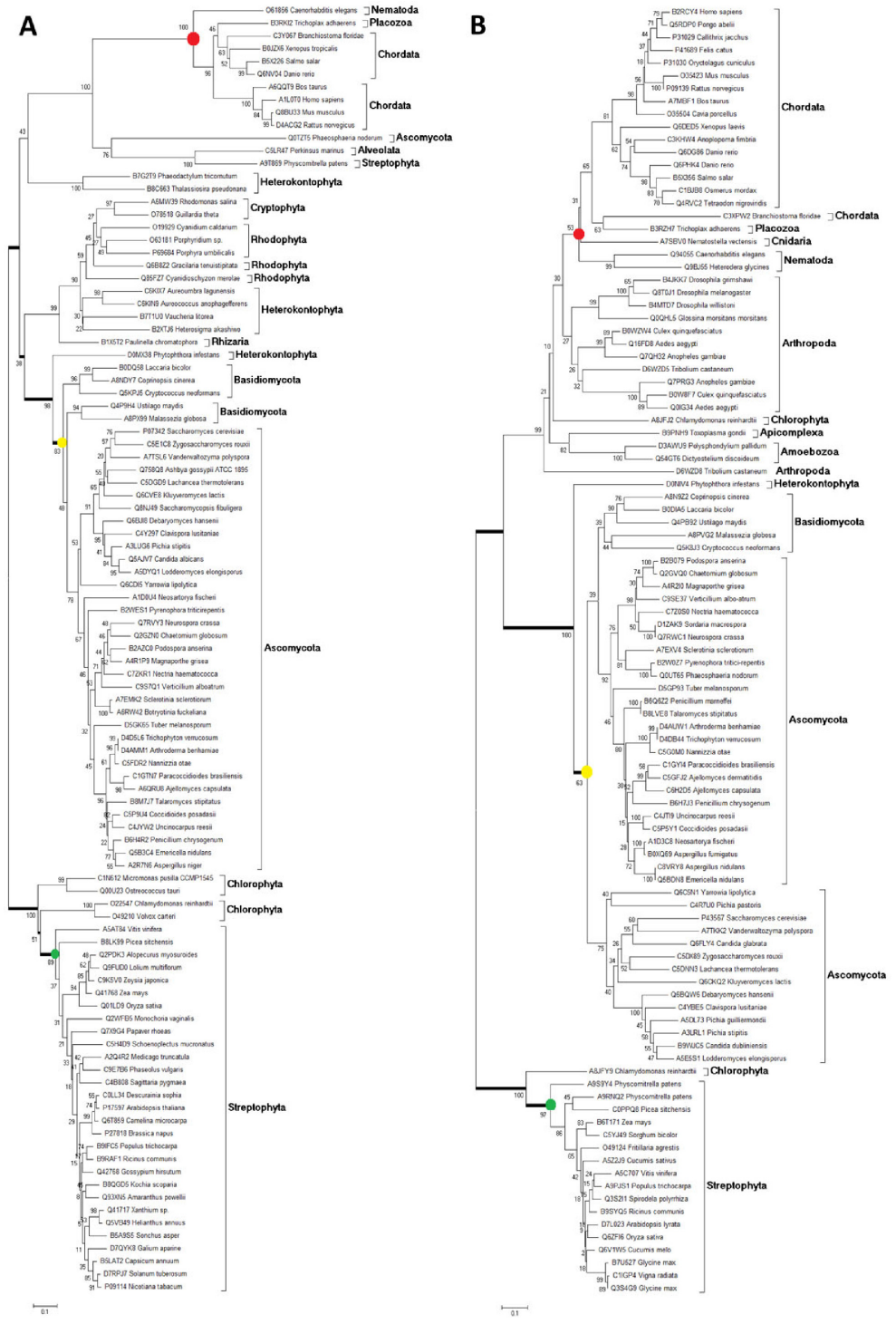


Figure 4 Phylogenetic analyses for EAA and NEAA enzymes. Phylogenetic trees for (A) acetolactate synthase (VIL1 code in Figure 1), an enzyme for EAA valine, isoleucine and leucine biosynthesis and (B) a group of alanine-glyoxylate, serine-glyoxylate and serine-pyruvate transaminases (G1 code in Figure 1), a NEAA biosynthetic enzyme for glycine biosynthesis. The green, yellow and red circles are marking the plant (Streptophyta), fungi (Dikarya) and animals (Metazoa) branches, respectively. In (A), the distance (given by substitutions per site) from the green circle to the yellow and red circles are, respectively, 0.4 and 1.0. In (B), these values are, respectively, 0.7 and 0.7.

experiencing higher divergence after the attainment of amino acids auxotrophy.

To support this observation, Figure 5 shows the ratios calculated for 12 enzymes. Only trees that show significant bootstraps for the branches of interest were considered. Enzyme codes in bars are described as in Figure 1. The Y axis at the right side corresponds to the distance measured from plant (Streptophyta) to the ancestor of fungi (Dikarya). This distance was assumed as a background distance to normalize the distances measured "from" plant (green bars) "to" the clades indicated in the X axis. The three enzymes on the right, S1, G1 and G2, belong to NEAA pathways, and the ratios are low. For the enzymes H5, FW7, F8, VIL1, VIL3, MT3 and M7, the ratio shown by green bars are conversely high, ranging from around 1.5 up to 7 fold. These preliminary data suggest that the additional evolutionary modifications have occurred in distinct levels in the enzymes maintained after the loss of biosynthetic capability. M(2) pathway appears as incomplete in Basidiomycota (Figure 1; M8 is absent), however MT3 enzyme used here is present in threonine pathway which is complete in this clade. K6 and K10 are involved in incomplete pathways, respectively, in plants and fungi. Accordingly, the distance measured from plant to fungi is high, and so is the drift between plant to Chordata (K6) or to Arthropoda (K10), therefore yielding balanced lower ratios. Since the ancestor of fungi and plants seems to be equally distant from both of these two groups, and the divergence between plant and Fungi/Metazoa group tends to a trifurcation (see Figure 4), the yellow bars (which represent the distance from fungi to the animal clades in the X axis divided by the background distance from plant to fungi) are similar to the ratios represented by the green bars, independently of how much modification has been occurred to the animal sequences (e.g. VIL1, MT3, G1). Furthermore, a detailed inspection of phylogenetic trees seems to indicate that subfunctionalized paralogues have appeared in basal clades such as Fungi, and those divergent paralogues remain in the more recent groups of organisms, while the copy that previously participated in the biosynthesis was actually deleted in animals. Note some Streptophyta and Ascomycota divergent paralogues (outparalogues) [27] grouped with animal sequences under 100% bootstrap (Figure 4A). Accordingly, similar divergent paralogues were observed for M7 enzyme (Ascomycota and Basidiomycota divergent paralogues grouped with animal sequences, 98% bootstrap, see additional file 2: Phylogenetic tree of 5-methyltetrahydropteroyltriglutamate–homocysteine methyltransferase (M7)). Moreover, for K10 enzyme that participates in the K biosynthetic pathway which is defective in fungi, a divergent paralogue from Streptophyta groups with fungi enzymes (92% bootstrap) near the Arthropoda sequence (Additional file 3: Phylogenetic tree of dihydrodipicolinate synthase (K10)). Thus, the enzymes remaining from biosynthetic pathways show higher divergence, and this might have been acquired due to subfunctionalization in ancient clades.

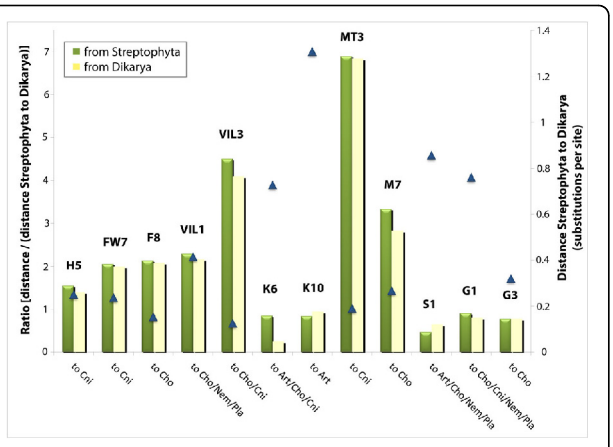


Figure 5 Relative distance of Metazoa enzymes from homologues of EAA and from NEAA biosynthetic enzymes present in plant and fungi. Phylogenetic trees were obtained for 12 enzymes, using all eukaryotic clustered proteins. Codes for enzymes are the same as in Figure 1 and are shown over the bars. For normalization, a background distance from the plant phylum Streptophyta to the fungi subkingdom Dikarya was measured and represented by triangles (right Y axis). The distance "from" either Streptophyta (green bars) or Dikarya (yellow bars), "to" the branches that group the clades indicated below the bars, were measured and normalized by the distance Streptophyta/Dikarya, yielding the ratio represented by bars (left Y axis). Only the three enzymes on the right (S1, G1 and G2) participate of biosynthesis of NEAAs: serine (S1) and glycine (G1 and G2). K6 and K10 are enzymes that compose lysine biosynthetic pathways which are not complete, respectively, in Streptophyta or Dikarya (see Figure 1). Abbreviations: Art, Arthropoda; Cho, Choanozoa; Cni, Cnidaria; Nem, Nematoda; Pla, Placozoa.

Discussion

The advance on genome sequencing and computational methods for clustering homologous proteins has been helping the scientific community to reevaluate several aspects of basic biology. Here we have applied clustering of protein sequences chosen from two clades of organisms that are known to be autotrophic for the biosynthesis of Essential Amino Acids (EAAs). Furthermore, we searched for the enzymes responsible for nitrogen assimilation, incorporating ammonium into glutamate. Lack of cytoplasmic glutamate dehydrogenase leads to a dependency of amino acids consumption as the source of organic nitrogen, i.e., the organism in a certain sense actually becomes auxotrophic to both EAAs and NEAAs (Non-Essential Amino Acids), in order to build other nitrogen-containing molecules.

The work presented here takes advantage of both the Seed Linkage software and a home-built UniProt Enriched KEGG Orthology database (UEKO) as source of information, to rapidly group homologues of fungi and plant amino acid sequences, respectively represented by *Saccharomyces cerevisiae* and *Arabidopsis thaliana*. KEGG Orthology contains to date more than 1 million sequences from nearly 1,000 genomes and it was enriched by a procedure developed by our group to attain 2,442,384 sequences from 25,024 organisms, constituting the UEKO database (UniRef50 enriched KEGG Orthology database, to be published elsewhere and further distributed). Counting the total recruited sequences reported in this work (31,392), the percentage of recruitment by (i) Seed Linkage, (ii) original KO or (iii) the enriched portion of KO (UEKO) was, respectively, 6%, 44% and 50%. Moreover, 26% of all detected enzymes for the phyla represented in Figures 1, 2 and 3 were exclusively detected by Seed Linkage software and/or UEKO database. These numbers reinforce the relevance on the development of homologous searching capability, improving the ability of KEGG Orthology database to build a scenario for the biological processes of interest such as those presented here. Moreover, on top of the search for homologues represented by circles in the Figures, a complementary search using the 31,392 clustered sequences allowed the investigation of all UniProt sequences, including fragments (e.g. UniProt accession B7QGP4, VIL1 from Arthropoda) and some full length proteins not accessed by the initial search (e.g. UniProt accession D3AYE6, complete protein K14, from Amoebozoa; actually a more recent version of KO already incorporates this entry). It is important to notice that, in UniProt, the technical term fragment is applied to partial CDS sequences, a product of incompletely sequenced mRNA, as well as amino acid sequences modeled from the genome that lack initial methionine. Thus they might represent additional evidence of the enzyme presence rather than a reminiscent pseudogene. Stringent criteria ($1x10^{-10}$ e-value, 50% identity and 50% subject coverage cutoffs) were adjusted with extensive manual inspection and additional evidences were included as triangles in the Figures. One evidence collected as triangle claimed our attention, since it came from a clade bearing the complete genome of the well annotated organism *Drosophila melanogaster* (Figure 1, enzyme VIL1, phylum Arthropoda). Manual inspection reveals that the evidence yielded by the additional search (represented by triangle) returned a hit from *Ixodes scapularis* (a genome under "assembly" status), but remarkably, the gene was found to be missing in the fly. Thus, this represents a recent gene loss within a non functional pathway.

The main interest of this work was to depict the evolution of amino acids essentiality, or heterotrophy. Grouping organisms into phyla level allowed easy labeling of clades that comprise organisms with sequenced or draft genomes, as shown in Figures 1, 2 and 3, making it possible to infer deletion events distinctively in these clades. It is important to notice that many phyla contain complete genomes, which allowed us to figure out the deletion process with more certainty. However, the picturing of the entire scenario allowed the analysis to be extended to the branched clades, although this requires additional caution on interpretation. Even escaping the scope of this work, it suggests a demand for planned choice of genomes to be completely sequenced, since as clearly shown here we lack information from several phyla such as the ones represented with empty circles (e.g. Cryptophyta, Haptophyta, Neocallimastigomycota and Glaucophyta). Enzymes not found by our analysis requires further attention and search using more sensitive methods and detailed manual or even experimental analysis, to detect divergent sequences; in other words, the absence of evidence is not evidence of absence. However, the present work exemplifies a method that can be easily applied to other scenarios of gene/pathway loss.

The scenario of amino acid auxotrophy supports the hypothesis of a Great Genomic Deletion model of amino acid biosynthesis in association with heterotrophy. This phenomenon has probably occurred several times, particularly at the origin of metazoans. This deletion has been likely associated with endosymbiotic relationships or with the development of systems specialized in nutrient absorption. It seems that amino acid essentiality has been originated as a phenotypic loss of pathways early in Choanozoa, followed by multiple losses during metazoan evolution. Similar progresses of deletions occur closer to Heterokontophyta and Rhizaria, culminating in Apicomplexa. Rhodophyta and Microsporidia also attain the auxotrophy.

Moreover, remaining enzymes set apart from their original roles in amino acid biosynthetic metabolism seem to be more prone to evolutionary changes whilst enzymes present in complete pathways are more structurally conserved among distant phyla (Figures 4 and 5). Although a detailed investigation is needed, our preliminary analysis suggests that the copies which remained in metazoan genomes may have suffered subfunctionalization and sometimes this might have occurred in more ancestral organisms (Figure 4 and additional files 2 and 3). Thus, in some sense, the orthologue enzyme might actually have been deleted in animals, and the divergent copy is the one remaining. These divergent copies are sometimes named outparalogues. We are currently

investigating substitution rate ratios and promoter elements in these genes.

Subsequent deletion includes the enzymes implicated in nitrogen assimilation, which takes place just after the broad deletion of EAAs biosynthetic enzymes (since except metazoans, other eukaryotic clades lack biosynthetic pathways and contains a nitrogen assimilative enzyme), as observed in more derived metazoans, but not Cnidaria. Most Cnidaria are carnivorous, so one possibility is that Cnidaria may benefit from the assimilation of organic nitrogen under long periods of fasting, however this finding needs additional investigation. Thus, the simplest explanation, is that the loss of nitrogen assimilative enzymes are related to lower selective pressure associated with the origin of the most heterotrophic organisms, animals.

To our knowledge this is the first initiative to clarify the complete scenario using powerful homologous grouping approaches and the total repertoire of sequenced genomes.

Conclusions

The procedures described here provide a deeper analysis of amino acid and nitrogen heterotrophy among distinct taxa, extended to include the entire set of available proteins. They show that amino acid essentiality was a broad phenomenon in eukaryotes, followed by the subsequent nutritional requirement of organic nitrogen, in animals.

Methods

Software and databases

Seed Linkage clustering software [14] and detailed explanation of usability can be obtained at http://www.biodados.icb.ufmg.br/eaa/. Seed Linkage requires BLAST (version used was 2.2.20), MySQL (version 5.0.77) [28] and PHP (version 5.1.6) [29].

The protein database is composed of UniProtKB entries (version used was 2010_09) available at http://www.biodados.icb.ufmg.br/eaa/. Except where otherwise indicated, all fragmented proteins were removed from analyses by parsing the description line in FASTA files.

To enrich KEGG Orthology clusters with incomplete genome proteins UniRef50 Enriched KEGG Orthology (UEKO) was built with the procedure described by Fernandes *et al* [15]. A local MySQL database was used.

Procedure

Amino acid biosynthetic pathways were depicted with KEGG Pathway [30] manual inspection where UniProtKB identifiers for the enzymes used in this work could be retrieved for the model autotrophic organisms *Saccharomyces cerevisiae*, *Arabidopsis thaliana* and, for the archaeal lysine biosynthesis, *Pyrococcus horikoshii*. The procedure starts with the selected sequences used as seed for Seed Linkage search in UniProtKB. The homologous cluster is enriched by (i) entries in KEGG Orthology (KO) belonging to the same KO where the seed is found and (ii) UEKO entries for this same KO. All steps were conducted with MySQL consults and PERL v5.8.8 [31] scripts. To verify the recruitment, seed sequences were used in PSI-BLAST alignments with the recruited sequences, having the PSI-BLAST iterations stopped whenever the score obtained for the seed sequence itself decreases to below 50% of the initial score. Results of search for homologues are represented by circles in the Figures. For more details see additional file 4: List of seed sequences and additional file 5: List of clusters.

Simple BLASTp analysis (10^{-10} e-value cutoff) were also conducted with all UniProt proteins, comprising both UniProt complete and fragment entries, for each phylum against all clustered proteins in this project. Resulting output was filtered to remove alignments with less than both 50% identity and 50% subject coverage. Results of this analysis are represented by triangles in the Figures.

Taxonomy information

All UniProtKB identifiers could be associated with an organism taxonomy ID with the file available at ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping.

Further association of organism taxonomy ID with phyla classification was achieved through a local database built with NCBI taxonomy information obtained at ftp://ftp.ncbi.nih.gov/pub/taxonomy.

Genome statuses were obtained by NCBI Genome Project analysis at: http://www.ncbi.nlm.nih.gov/genomeprj.

Phylogenetic analyses

For phylogenetic analysis Prankster [32] was used for multiple sequence alignment and MEGA4 [33] to construct the phylogenetic tree using the neighbor-joining method [34] with 500 bootstrap replicates. Branch distances were obtained from phylogenetic trees, from the ancestors of Streptophyta, Dikarya and clades of metazoans. Only branches with significant bootstrap were used. With the distances, a ratio was calculated as below:

$$\textbf{Distance F - T / Distance S - D}$$

where F (from) is either Streptophyta or Dikarya ancestor and T (to) is an animal ancestor (see Figure 5, X axis); and S and D are the ancestors of Streptophyta and Dikarya, respectively. Phylogenetic trees used to compose Figure 5 can be accessed at our server at http://www.biodados.icb.ufmg.br/eaa/.

Additional material

Additional file 1: Sequences and genome status distribution. Distribution of UniProtKB sequences among available genomes in three sequencing status groups: Complete, Draft plus In Progress and Incomplete.

Additional file 2: Phylogenetic tree of 5-methyltetrahydropteroyltriglutamate–homocysteine methyltransferase (M7). A phylogenetic tree of one of the four methyltransferases illustrated in Figure 1 for methionine biosynthesis. Red circle represents Chordata and Cnidaria ancestor; Yellow circle Dikarya ancestor and green circle Streptophyta ancestor. Available at [http://www.biodados.icb.ufmg.br/eaa/].

Additional file 3: Phylogenetic tree of dihydrodipicolinate synthase (K10). A phylogenetic tree of one of the enzymes illustrated in Figure 1 for lysine biosynthesis. Red circle represents Arthropoda; Yellow circle Dikarya ancestor and green circle Streptophyta and Chlorophyta ancestor. Available at [http://www.biodados.icb.ufmg.br/eaa/].

Additional file 4: List of seed sequences. A detailed list of sequences used as initiators for clustering process with UniProtKB identifier, NCBI taxonomy identifier and Enzyme Commission (EC) number. Available at [http://www.biodados.icb.ufmg.br/eaa/].

Additional file 5: List of clusters. A detailed list of created clusters for all enzymes with UniProtKB identifier and NCBI taxonomy identifier. Available at [http://www.biodados.icb.ufmg.br/eaa/].

List of abbreviations

COG: Cluster of Orthologous Groups; EAAs: Essential Amino Acids; GDH: Glutamate dehydrogenase; KEGG: Kyoto Encyclopedia of Genes and Genomes; KO: KEGG Orthology; NEAAs: Non-Essential Amino Acids; UEKO: UniRef50 Enriched KEGG Orthology.

Acknowledgements

Authors thank Dr. Darren Natale from PIR (USA) and Elisa Donnard (LICR) for critically reviewing this manuscript, Henrique Velloso for helping with taxonomic data and Laryssa Santos Queiroz with pathway inspections. This work has been sponsored by the Brazilian Ministry of Education (CAPES) and Foundation for Research Support of Minas Gerais State (FAPEMIG).
This article has been published as part of *BMC Genomics* Volume 12 Supplement 4, 2011: Proceedings of the 6th International Conference of the Brazilian Association for Bioinformatics and Computational Biology (X-meeting 2010). The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2164/12?issue=S4

Author details

[1]Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, 31270-901, MG, Brazil. [2]Programa de Pós-Graduação em Ciências Genômicas e Biotecnologia, Universidade Católica de Brasilia, Brasilia, 70790-160, DF, Brazil. [3]Instituto de Bioquímica Médica, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 21941-902, RJ, Brazil.

Authors' contributions

The work presented here was carried out in collaboration between all authors. FP and JMO defined the research theme. RLMG developed the clustering procedure, created the dataset and conducted the experiments. RLMG and GFR created the figures. RLMG, FP and LKM conducted phylogenetic analyses. GRF created the procedure of Uniref50 enrichment of KEGG Orthology database. HALR developed the PSI-BLAST validation method. JMO, FP and RLMG wrote the paper. All authors supervised and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Published: 22 December 2011

References

1. Blair J, Shah P, Hedges SB: **Evolutionary sequence analysis of complete eukaryote genomes.** *BMC Bioinformatics* 2005, **6**:53.
2. Cunchillos C, Lecointre G: **Early steps of metabolism evolution inferred by cladistic analysis of amino acid catabolic pathways.** *Comptes Rendus Biologies* 2002, **325**:119-129.
3. Cunchillos C, Lecointre G: **Ordering events of biochemical evolution.** *Biochimie* 2007, **89**:555-573.
4. Hernández-Montes G, Díaz-Mejía JJ, Pérez-Rueda E, Segovia L: **The hidden universal distribution of amino acid biosynthetic networks: a genomic perspective on their origins and evolution.** *Genome Biol* 2008, **9**:R95.
5. Reeds PJ, Wahle KWJ, Haggarty P: **Energy costs of protein and fatty acid synthesis.** *Proceedings of the Nutrition Society* 1982, **41**:155-159.
6. Aoyagi Y, Tasaki I, Okumura J, Muramatsu T: **Energy cost of whole-body protein synthesis measured in vivo in chicks.** *Comp Biochem Physiol A Comp Physiol* 1988, **91**:765-768.
7. Millward DJ: **Metabolic Demands for Amino Acids and the Human Dietary Requirement: Millward and Rivers (1988) Revisited.** *The Journal of Nutrition* 1998, **128**:2563S-2576S.
8. Millward DJ, Rivers JP: **The nutritional role of indispensable amino acids and the metabolic basis for their requirements.** *Eur J Clin Nutr* 1988, **42**:367-393.
9. Elango R, Ball R, Pencharz P: **Amino acid requirements in humans: with a special emphasis on the metabolic availability of amino acids.** *Amino Acids* 2009, **37**:19-27.
10. Payne SH, Loomis WF: **Retention and Loss of Amino Acid Biosynthetic Pathways Based on Analysis of Whole-Genome Sequences.** *Eukaryotic Cell* 2006, **5**:272-276.
11. Consortium TU: **The Universal Protein Resource (UniProt) in 2010.** *Nucleic Acids Research* 2010, **38**:D142-D148.
12. Tatusov R, Fedorova N, Jackson J, Jacobs A, Kiryutin B, Koonin E, Krylov D, Mazumder R, Mekhedov S, Nikolskaya A, *et al*: **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics* 2003, **4**:41.
13. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M: **The KEGG resource for deciphering the genome.** *Nucleic Acids Research* 2004, **32**:D277-D280.
14. Barbosa-Silva A, Satagopam V, Schneider R, Ortega JM: **Clustering of cognate proteins among distinct proteomes derived from multiple links to a single seed sequence.** *BMC Bioinformatics* 2008, **9**:141.
15. Fernandes GR, Barbosa DVC, Prosdocimi F, Pena IA, Santana-Santos L, Coelho Junior O, Barbosa-Silva A, Velloso HM, Mudado MA, Natale DA, *et al*: **A procedure to recruit members to enlarge protein family databases–the building of UECOG (UniRef-Enriched COG Database) as a model.** *Genetics and molecular research GMR* 2008, **7**:910-924.
16. Suzek B, Huang H, McGarvey P, Mazumder R, Wu C: **UniRef: comprehensive and non-redundant UniProt reference clusters.** *Bioinformatics* 2007, **23**:1282-1288.
17. Katinka MD, Duprat S, Cornillot E, Metenier G, Thomarat F, Prensier G, Barbe V, Peyretaillade E, Brottier P, Wincker P, *et al*: **Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi.** *Nature* 2001, **414**:450-453.
18. Prosdocimi F, Mudado MA, Ortega JM: **A set of amino acids found to occur more frequently in human and fly than in plant and yeast proteomes consists of non-essential amino acids.** *Computers in Biology and Medicine* 2007, **37**:159-165.
19. Santana-Santos L, Prosdocimi F, Ortega JM: **Essential amino acid usage and evolutionary nutrigenomics of eukaryotes–insights into the differential usage of amino acids in protein domains and extra-domains.** *Genetics and molecular research GMR* 2008, **7**:839-852.
20. Miyazaki T, Miyazaki J, Yamane H, Nishiyama M: **α-Aminoadipate aminotransferase from an extremely thermophilic bacterium, Thermus thermophilus.** *Microbiology* 2004, **150**:2327-2334.
21. Velasco AM, Leguina JI, Lazcano A: **Molecular Evolution of the Lysine Biosynthetic Pathways.** *Journal of Molecular Evolution* 2002, **55**:445-449.
22. Hudson AO, Bless C, Macedo P, Chatterjee SP, Singh BK, Gilvarg C, Leustek T: **Biosynthesis of lysine in plants: evidence for a variant of the known bacterial pathways.** *Biochim Biophys Acta* 2005, **1721**:27-36.
23. Nishida H, Nishiyama M, Kobashi N, Kosuge T, Hoshino T, Yamane H: **A Prokaryotic Gene Cluster Involved in Synthesis of Lysine through the Amino Adipate Pathway: A Key to the Evolution of Amino Acid Biosynthesis.** *Genome Research* 1999, **9**:1175-1183.

24. Smith TJ, Schmidt T, Fang J, Wu J, Siuzdak G, Stanley CA: **The Structure of Apo Human Glutamate Dehydrogenase Details Subunit Communication and Allostery.** *Journal of Molecular Biology* 2002, **318**:765-777.
25. Allen A, Kwagh J, Fang J, Stanley CA, Smith TJ: **Evolution of Glutamate Dehydrogenase Regulation of Insulin Homeostasis Is an Example of Molecular Exaptation†.** *Biochemistry* 2004, **43**:14431-14443.
26. Andersson J, Roger A: **Evolution of glutamate dehydrogenase genes: evidence for lateral gene transfer within and between prokaryotes and eukaryotes.** *BMC Evolutionary Biology* 2003, **3**:14.
27. Sonnhammer ELL, Koonin EV: **Orthology, paralogy and proposed classification for paralog subtypes.** *Trends in Genetics* 2002, **18**:619-620.
28. **MySQL.** [http://www.mysql.com].
29. **PHP.** [http://www.php.net].
30. Kanehisa M: **KEGG: From genes to biochemical pathways.** *Bioinfomatics: Databases and Systems* Kluwer Academic Publishers; 1999, 63-76.
31. **Perl.** [http://www.perl.org/].
32. Löytynoja A, Goldman N: **An algorithm for progressive multiple alignment of sequences with insertions.** *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**:10557-10562.
33. Tamura K, Dudley J, Nei M, Kumar S: **MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0.** *Molecular Biology and Evolution* 2007, **24**:1596-1599.
34. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Molecular Biology and Evolution* 1987, **4**:406-425.

doi:10.1186/1471-2164-12-S4-S2
Cite this article as: Guedes *et al.*: Amino acids biosynthesis and nitrogen assimilation pathways: a great genomic deletion during eukaryotes evolution. *BMC Genomics* 2011 **12**(Suppl 4):S2.

12. Anexo II

Donnard et al. BMC Genomics 2011, 12(Suppl 4):S3
http://www.biomedcentral.com/1471-2164/12/S4/S3

BMC Genomics

PROCEEDINGS Open Access

Preimplantation development regulatory pathway construction through a text-mining approach

Elisa Donnard[1,2], Adriano Barbosa-Silva[3,4], Rafael LM Guedes[1], Gabriel R Fernandes[1], Henrique Velloso[1], Matthew J Kohn[5], Miguel A Andrade-Navarro[3], J Miguel Ortega[1*]

From 6th International Conference of the Brazilian Association for Bioinformatics and Computational Biology (X-meeting 2010)
Ouro Preto, Brazil. 15-18 November 2010

Abstract

Background: The integration of sequencing and gene interaction data and subsequent generation of pathways and networks contained in databases such as KEGG Pathway is essential for the comprehension of complex biological processes. We noticed the absence of a chart or pathway describing the well-studied preimplantation development stages; furthermore, not all genes involved in the process have entries in KEGG Orthology, important information for knowledge application with relation to other organisms.

Results: In this work we sought to develop the regulatory pathway for the preimplantation development stage using text-mining tools such as Medline Ranker and PESCADOR to reveal biointeractions among the genes involved in this process. The genes present in the resulting pathway were also used as seeds for software developed by our group called SeedServer to create clusters of homologous genes. These homologues allowed the determination of the last common ancestor for each gene and revealed that the preimplantation development pathway consists of a conserved ancient core of genes with the addition of modern elements.

Conclusions: The generation of regulatory pathways through text-mining tools allows the integration of data generated by several studies for a more complete visualization of complex biological processes. Using the genes in this pathway as "seeds" for the generation of clusters of homologues, the pathway can be visualized for other organisms. The clustering of homologous genes together with determination of the ancestry leads to a better understanding of the evolution of such process.

Background

Bioinformatics tools currently allow research to focus on the integration of large-scale data generated by sequencing, differential expression analysis, gene interaction studies and others. Several initiatives exist to organize this knowledge in secondary databases, thus allowing easier access and visualization. Databases containing interaction information are a good source for novel research. iHOP [1] allows users to tag gene names of interest and browse through the related PubMed literature with highlighted keywords. Another interaction database is STRING [2], which contains physical interactions and functional associations between proteins and integrates data retrieved from literature (PubMed), genomic context, large scale experiments and conserved co-expression. Text-mining, therefore, has a fundamental role in these tools and allows access to interactions spread throughout the literature. The extraction of biological events from literature through text-mining tools is essential to not only update the interaction databases but also for the creation and annotation of pathways.

Metabolic and regulatory pathways are an example of organized knowledge that allow a better visualization of a complex system and can be found in databases such as iPath [3], BioCyc [4] or KEGG Pathways [5]. When orthology information is added to pathways, the same

* Correspondence: miguel@ufmg.br
[1]Laboratório Biodados, Dept. de Bioquímica e Imunologia, Universidade Federal de Minas Gerais, 31270-901, Belo Horizonte - MG, Brazil
Full list of author information is available at the end of the article

process can be represented in different organisms. Orthology is also an important tool for sequence annotation. Current orthologue databases such as COG and KOG [6], eggNOG [7], OrthoMCL [8] and KEGG Orthology [5] all provide a good source for manually curated clusters of orthologues defined for organisms with complete genomes. We developed a procedure to enrich the COG database with UniRef50 clusters from the UniProt database [9], creating the UECOG database [10]. Recently, a similar procedure was applied to the KEGG Orthology database creating the enriched UEKO database (unpublished, Fernandes *et al.*).

The available tools described raise the possibility of integrating current information and generating complex regulatory pathways. Previous publications individually reported the regulatory interactions that control preimplantation embryo development [11-15]. However, a complete preimplantation development regulatory pathway has never been built.

In humans, the preimplantation phase of embryonic development is a period of approximately six days after fertilization prior to attachment of the embryo to the uterine wall. Implantation can occur before or in the seventh embryonic day (E7), a time during which the uterus is receptive [16]. Mammalian embryonic development has been thoroughly studied in mice and the blastomeres remain totipotent, able to generate any other cell, up to the eight-cell stage, unlike other animals [17]. After fertilization, successive cleavages take place during the first two days of development, resulting in the eight-cell embryo. The next stage of development is called the morula stage. An increase in cell-cell contact results in formation of a compacted morula. The subsequent divisions increase the complexity of the embryo and cells may be located on the inside, surrounded by other cells, or on the outside, in contact with the environment. The identification of the initial cells for each lineage has shown that the trophectoderm (TE) is derived mostly from the outer cells, whereas the inner cells give rise to the inner cell mass (ICM). Later, the ICM divides into the primitive endoderm (PE) and the epiblast (EPI). During the differentiation of the TE from the ICM, the blastocoel is formed through a process of cavitation. The embryo is called a blastocyst when all three structures are present (TE, ICM and blastocoel). Twenty-four hours after blastocyst formation occurs, the last stage of preimplantation development takes place when the PE differentiates from the ICM. The three lineages thus formed in preimplantation development present different fates during subsequent embryonic development. While the epiblast, which forms from the ICM following implantation, is still undifferentiated and will give rise to the fetus itself, the trophectoderm will become the fetal portion of the placenta and the primitive endoderm (as part of the extraembryonic endoderm) will form the yolk sac [14]. Complex regulatory processes such as animal development are a result of the interaction of many different gene products and elements that control the expression of these genes. Traditional experiments that determine the function of one or a few genes are essential, but do not result in a comprehensive view of complex systems. A complex regulatory network should be able to portray specific and general aspects of development, such as the embryonic fate of certain cells [18].

In this work, we noticed the absence in databases of a pathway describing the preimplantation phase of embryo development and sought to develop the given pathway using text-mining tools, complementing it with orthology information. The resulting pathway comprises 86 genes and the interactions between them. Clusters of orthologous groups were generated for each gene represented in the pathway and provided the necessary information to determine the last common ancestor. This determination revealed that the preimplantation development pathway is an ancient Chordata pathway with addition of modern elements throughout evolution.

Results

Text-mining

Initially, we used the PubMed platform to search for articles related to the embryo preimplantation development (query: "preimplantation development") and obtained 3524 entries as a result. To obtain a more efficient set of articles with relevancy to our work, the result entries were submitted to MedlineRanker [19]. This software computes discriminating words by comparing a set of user selected abstracts indicated as highly relevant to a background set and then scores any abstracts in terms of their content of those discriminating words. After the classification, we selected the top 1000 abstracts for further analysis, which presented p-value lower than 0.01 and by manual inspection provide large amount of information when uploaded in PESCADOR. Since human and mouse embryo development are highly similar, it was plausible to use abstracts from work on both organisms as source of information for the preimplantation pathway construction, paying attention to any possible conflict.

Using these 1000 highly informative abstracts as our input, PESCADOR (manuscript in preparation, Barbosa-Silva *et al.*) an online platform for friendly operation of the LAITOR software [20]) was used for tagging of gene names and biointeractions extraction from each abstract. As a result, 722 gene names were tagged and 223 type 1 biointeractions were highlighted as well as other informative biointeractions. Biointeractions are classified by LAITOR [20] as type 1 when in the same sentence the software encounters a gene name, a biointeraction word

and another gene name, in that order (e.g.: CDX2 down-regulates NANOG). From these tagged abstracts we manually curated the information and constructed the pathway for the preimplantation embryo development describing 86 genes and numerous interactions between them during the early developmental stages, trophectoderm differentiation from the inner cell mass and posterior extraembrionary endoderm differentiation. A sample abstract tagged by PESCADOR and the manual extraction of the information it contains is exemplified in Figure 1. The pathway shown in Figure 2 was constructed according to KGML (KEGG Markup Language). The large decrease in the initial number of genes tagged in the abstracts is mainly due to redundancy between abstracts (same genes mentioned) and also to genes tagged in type 3 and 4 biointeractions, which not always result in pathway building information.

Preimplantation pathway

The pathway obtained after the analysis of all the abstracts from the PESCADOR output is represented in Figure 2 and the regulations are reviewed below.

First embryonic cleavages

The oncogene c-MYC is an important transcriptional regulator and its expression is observed in the initial stages of development, where it is present in embryonic cells until the morula stage and repressed thereafter [21]. Two additional genes recently associated with these early developmental stages are BORIS and ECSA. BORIS is involved in early development following fertilization and soon afterwards repressed, and ECSA, expression begins in the blastocyst exclusively in the cells of the inner cell mass (ICM). The presence of these genes was compared to the expression pattern of the Oct4 transcription factor, which is present in the early cleavages, repressed after this initial stage, and

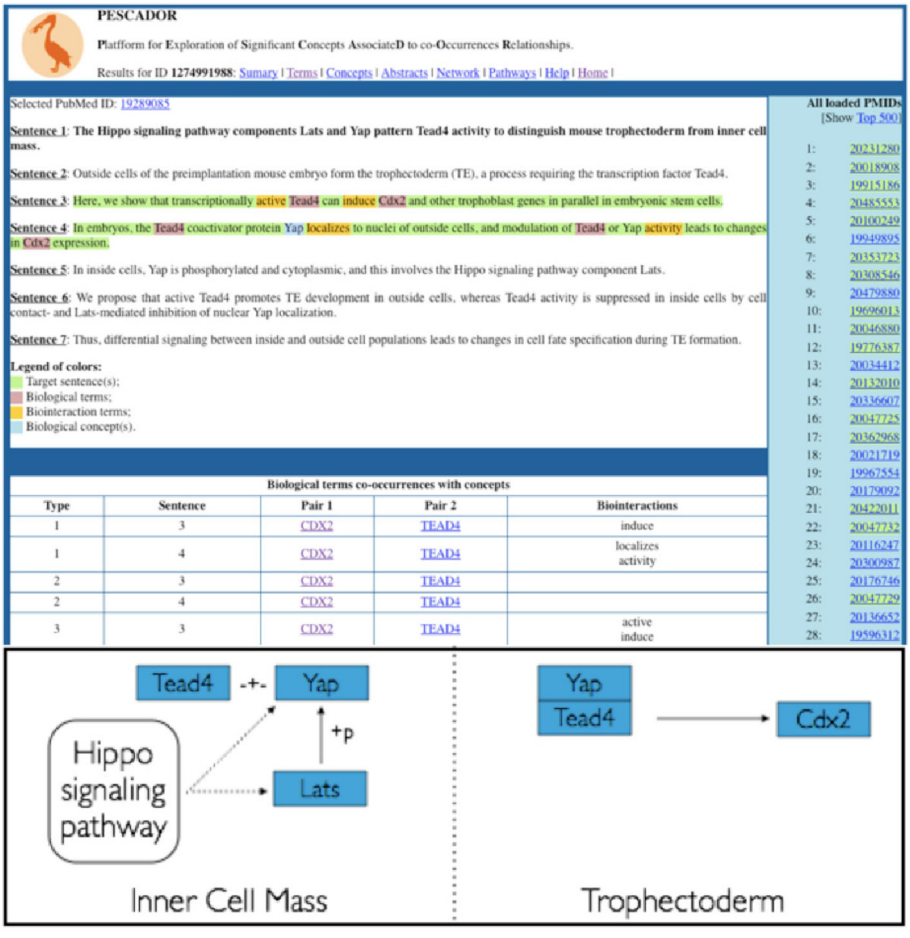


Figure 1 Biointeraction extraction from PESCADOR. Top: Sample abstract tagged by PESCADOR. Gene or protein names (terms) recognized are highlighted in violet and the biointeraction words in yellow. The platform allows users to search for their interactions of interest by terms, abstracts or concepts of interest added initially by the user. Bottom: Manual curation of the information presented in the abstract and its graphical representation in the form of a regulatory pathway.

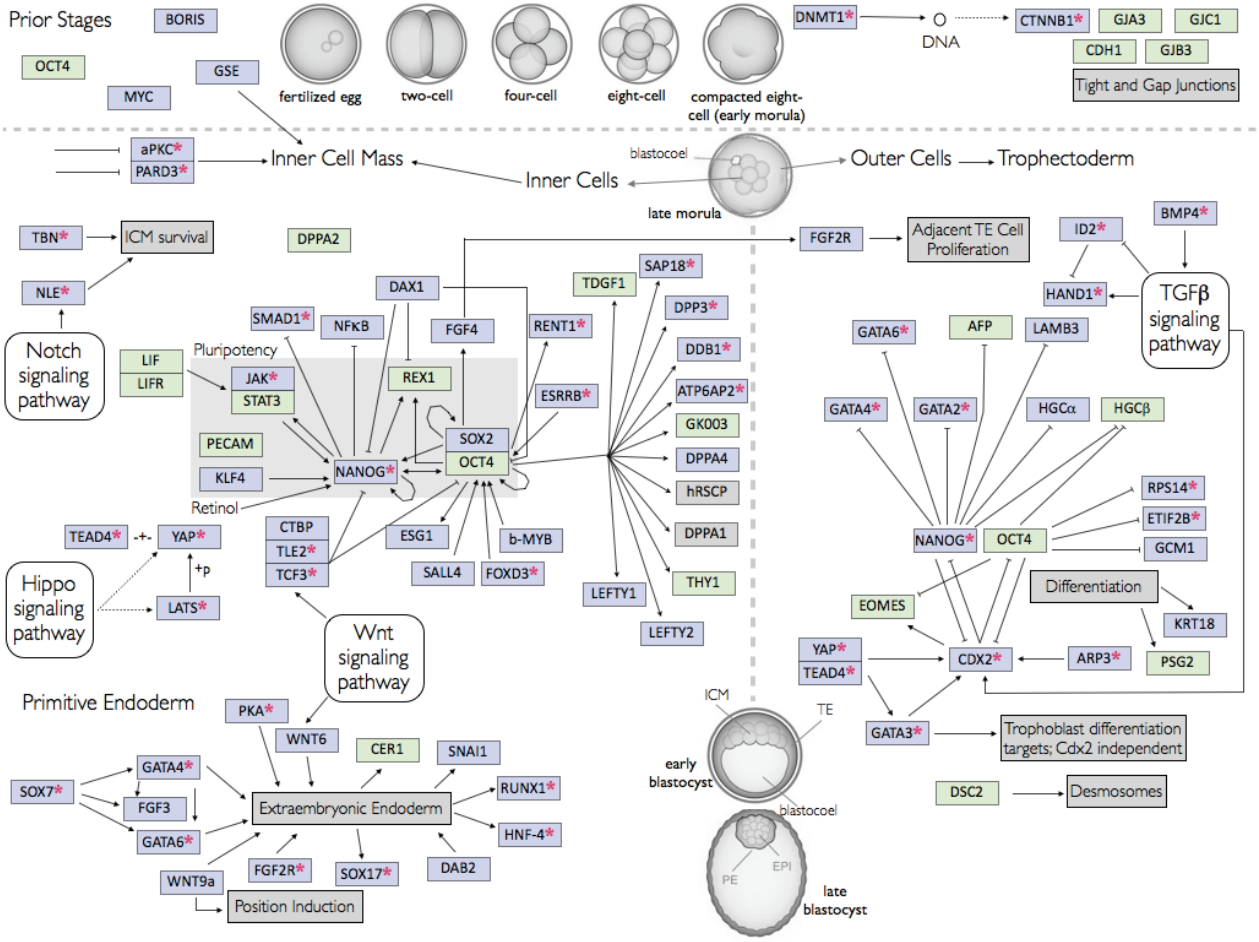


Figure 2 Preimplantation development pathway. The figure shows a pathway representation of the genes involved in the regulation of the preimplantation development and interactions between them. Some functions are also detailed in the grey rectangles. In the upper part of the figure are located genes involved in the early stages of development (until blastocyst formation) below these, the left part corresponds to the regulations that occur in the inner cell mass, the portion of cells that remains undifferentiated during a longer period of time, part of these cells will give rise to the primitive endoderm and the genes that regulate this process are shown in the bottom left. In the right are the genes involved in the development of the outer cells of the blastocyst, which differentiate to form the trophectoderm. The interactions are described in the text. KEGG Markup Language was used for pathway representation. The developmental stages figures were adapted from Yamanaka *et al.* 2006 [14]. The pathway genes are represented according to their ancestry based on the determination of their Last Common Ancestor. Genes considered recent are shown in green while genes of more ancient origin are shown in lilac. Genes that present an ortholog in *D. melanogaster* are marked (*). This will be further adressed in the text section "Pathway Ancestry". DPPA1 and hRSCP are shown in grey due to the fact that the lack of corresponding SwissProt annotated gene product to be used as seed prevented their use in this analysis.

then its expression is afterwards stimulated again in the blastocyst [22]. The expression of the gametogenesis associated gene Gse was also recently identified in cells of the early embryo; later this protein is found only in the ICM, suggesting a role in the specification of cell lineage [23].

Methylation patterns and correct preimplantation development

Genomic methylation patterns in mammalian cells depend on Dnmt1 (DNA methyltransferase-1). In the mouse, an embryo-specific variant called Dnmt1o is expressed in the early stages of development. In the 8-cell stage this protein relocates to the cell nucleus where maintains essential methylation patterns, allowing embryos to complete early developmental events [24]. It was recently shown [25] that the inability of Dnmt1o to properly relocate not only results in a developmental arrest at the 5-7 cell stage, but is also responsible for the downregulation of five genes involved in the formation of gap and tight junctions (Cx31, Cx43, Cx45, Cdh1 and Ctnnb1). These junctions are crucial for early processes such as compaction of the 8-cell embryo and cavitation of the blastocoel.

TE versus ICM dichotomy: key role of Lats controlling Tead4 co-activator Yap

Cells destined to become part of the ICM are marked by repression of two genes (aPKC and PARD3) [26] and by upregulation of Sox2 [11]. In these cells, the major pluripotency transcription factors, including Nanog and Oct4, remain active due to the expression of an important player and member of the Hippo signaling pathway: Lats. This serine/threonine protein kinase is responsible for phosphorylating Yap, leading to its cytoplasmic localization and thus preventing its association with the transcription factor Tead4.

Triggering TE differentiation: Tead4/Yap target Cdx2 to repress Nanog and Oct4

Conversely, in the outer cells that will differentiate and form the trophectoderm, Yap is unphosphorylated, remains in the nucleus and associates with Tead4, leading to the activation of Cdx2, a key repressor of Nanog and Oct4 [27]. Repression of Oct4 and Nanog transcription by Cdx2 then releases the inhibition that these two key factors were exerting on many different genes, in turn activating these targets [28,29]. Activation of Cdx2 requires release from basal repression; Nanog [30] and Oct4 [31] repress basal levels of Cdx2 and induction of higher levels of Cdx2 by Tead4/Yap overcomes this repression, allowing Cdx2 to play its role [28]. Tead4 was also recently determined to activate another trophectoderm differentiation factor, GATA3 [32], which acts alongside Cdx2 and affects transcription of a number of genes independent of Cdx2. The Tead4-dependent activation of GATA3 seems to be independent of Yap, suggesting Tead4 interacts with another partner as well as Yap. Also required for high level expression of Cdx2 in trophectoderm cells is the cell motility protein Arp3; experiments with complete knockdown of this protein show trophoblast cells unable to develop properly, possibly undergoing apoptosis as a result of loss of Cdx2 [33]. The TGFbeta pathway is another important pathway for trophectoderm differentiation; TGFbeta signaling is stimulated by BMP4, which leads to the activation of SMAD proteins. These proteins can also stimulate transcription of Cdx2 [34], and BMP4 is known to inhibit Id2, an inhibitor of differentiation [35], and to activate Hand1, which is involved in trophoblast cell differentiation [36].

In the absence of Oct4 and Nanog

The downregulation of Oct4 in the outer cells of the embryo leads to the activation of a positive regulator of TE cell fate, Eomes (T-box protein eomesodermin) [29,37], which is also a possible Cdx2 target [38]. The subsequent differentiation of these cells into trophectoderm is accompanied by the expression of several genes, such as the glycoprotein PSG2 [39] and the marker KRT18. PSG2 and KRT18 expression are among the first signs that a blastomere has lost its totipotent competence, prior to any visible differentiation [33]. Removal of Oct4-dependent repression also results in activation of genes such as ETIF2B and Rps14 [40], allowing these cells to engage in an intense translation routine. Knockdown studies targeting Oct4 also show that it represses the expression of Gcm1, which is normally placenta specific [41], and of the hCG hormone's beta chain [42].

Concurrently, Nanog downregulation allows the expression of a number of genes associated with both trophectoderm (GATA2, hCG-alpha and hCG-beta) and extraembryonic endoderm (GATA4, GATA6, LAMB1 and AFP) [30]. These latter genes will in turn initiate the formation of tissues such as the primitive endoderm, a component of the yolk sac. From the early blastocyst stage on, desmosomes are assembled in the trophectoderm in response to desmocollin (DSC2), which is also not expressed in the ICM [43].

Thus, Tead4/Yap activation of Cdx2, accompanied by the subsequent repression of Nanog and Oct4, describes a scenario for the TE differentiation.

Underneath the maintained activation of Oct4 and Nanog

Back in the ICM, the main pluripotency genes remain active and form a complex regulation pathway. Recently it was discovered that transcription of Nanog is further stimulated by the presence of compounds such as retinol [44]. Klf2, Klf4 and Klf5 exert a redundant role in the activation of Nanog. These krüppel-like factors were described as essential for the maintenance of pluripotency. Indeed, Klf4 was already known for this role and is commonly used in reprogramming of differentiated cells into induced pluripotent stem cells. However, only the simultaneous depletion of Klf4, 2 and 5 results in the differentiation of stem cells, indicating functional redundancy [45]. Other proteins known to activate Nanog include the two other main pluripotency regulators, Oct4 [37,46] and Sox2 [47]. The estrogen receptor ESRRB is also reported to be involved in the activation of Nanog by Oct4 and Sox2 [47]. Conversely, Nanog can activate Oct4 [46], and ESRRB is necessary to maintain Oct4 promoter activity [48].

Each of the three key factors, Oct4, Sox2 and Nanog, also act as self-activators, e.g. the partners Oct4 and Sox2 bind and activate Oct4 transcription [49]. Another key transcription factor involved in the maintenance of cell pluripotency is Sall4 [50]. Sall4 binds to the conserved regulatory region in the Pou5f1 (the Oct4 gene) distal enhancer and activates its transcription [31]. Studies with microRNA interference of Sall4 show that the loss of this factor leads to reduction of Oct4 mRNA levels and significant expression of Cdx2 in the ICM [31]. b-MYB, a gene expressed in proliferating cells, is also a positive regulator of Oct4 and studies report early differentiation of ICM in the absence of b-MYB [51].

The Notch signaling pathway is a conserved pathway that is involved in cellular communication processes and correct cell fate decisions that also has a role in ICM development [52]. Nle protein, a direct regulator of this pathway, is essential for survival of the ICM [53]. Another protein associated with development and survival of the ICM is Tbn (Taube nuss), whose absence promotes cell apoptosis in the ICM [54].

Expression of the platelet and endothelial cell adhesion molecule (PECAM1 or CD31) was detected by immunofluorescence confocal microscopy in the blastocyst and restricted to the ICM cells. Subsequently, PECAM1 remains only in the pluripotent epiblast cells, disappearing the moment these cells undergo differentiation [55], and indicating a new role for this molecule during embryo development.

Activation, but with moderation

Other control pathways maintain expression of these genes at a steady-state concentration and balance these many mechanisms for activation and upregulation of transcription. A complex regulation feedback loop consists of FOXD3, Nanog and Oct4 [46]. To keep Oct4 and Nanog expression within steady-state levels, these three genes interact so that (i) expression of Nanog activates FOXD3 and Oct4 but not above steady-state levels due to Oct4 exerted repression; and (ii) FOXD3 and Nanog activate Oct4 expression but not above steady-state levels due to Oct4 self-repression.

Dax1 is an orphan nuclear hormone receptor recently identified as a repressor of Oct4 transcription [56].

Dax1 expression was also capable of reducing Nanog and Rex1 expression. Assays show that Dax1 binds to Oct4 and abolishes its DNA binding activity, thus decreasing the transcription of Nanog and Rex1, targets of Oct4 activation.

Another repressor in the ICM is Tcf3, a Wnt signaling pathway effector. TLE2 (a Groucho family protein) and CtBP (C-terminal binding protein) are key partners of Tcf3 in mediating this repressive effect. Tcf3 binds to and represses the Oct4 promoter, and this repressive effect requires both the Groucho and CtBP interacting domains of Tcf3 [13]. Tcf3 also limits the steady-state levels of Nanog mRNA, protein, and promoter activity in self-renewing embryonic stem cells (ESCs); the Tcf3 Groucho domain is involved in this repression [57]. Thus, Tcf3 is critical for maintaining the appropriate levels of both Oct4 and Nanog in ESCs. Experiments show that loss of Tcf3 by RNA interference (RNAi) knockdown blocks the ability of ESCs to differentiate [13], emphasizing the importance of this interaction.

Downstream of Oct4 and Sox2

Oct4 activates embryonic stem cell-specific gene 1 (Esg1), which encodes an RNA binding protein present in the ICM that is responsible for regulating several specific target transcripts [58]. Oct4 and Sox2 are also responsible for the regulation of the fibroblast growth factor 4 (FGF4) [49]. Expression of FGF4, therefore, requires the combined activity of these two transcription factors that bind to adjacent sites on the FGF4 enhancer DNA region [59]. Once expressed, the FGF4 protein can interact with its receptor FGFR2 and activate ICM and adjacent TE cell proliferation, activating extraembryonic endoderm cells as well in later stages.

Several other genes with important functions in embryonic development are also targets of Oct4-dependent activation. These include growth factor TDGF1, growth inhibitor SAP18, regulator of nonsense transcripts RENT1, two proteins involved in stem cell self-renewal DPPA4 and DPPA1 (developmental pluripotency associated), anterior visceral endoderm (AVE) markers LEFTY1 and LEFTY2, surface antigen THY1, and other genes encoding proteins involved in specialized cellular processes (DPP3, ATP6AP2, DDB1) and hypothetical proteins (GK003, hRscp) [37,40].

The master regulation exerted by Sox2 and Oct4 during mammalian embryogenesis is believed to operate through their cooperative binding to DNA regulatory regions composed of adjacent HMG and POU motifs (HMG/POU cassettes) [60]. Exemplifying this arrangement, DPPA4 is one such gene with the presence of an HMG/POU cassette in its promoter region [61].

Downstream of Nanog and STAT3

Activation of JAK/STAT pathway also has an important contribution to pluripotency. In mice, the LIF/STAT3 pathway [44,62,63] for maintenance of cell pluripotency comprises LIF and LIF receptor, which deliver intracellular signaling through STAT3. STAT3, a signal transducer and activator of transcription is activated by the JAK1 kinase and binds to several promoters inducing transcription of pluripotency related genes [64]. Nanog and Stat3 were found to bind to and synergistically activate Stat3-dependent promoters [64]. Nanog also functions as a transcriptional inhibitor to NFκB, a factor known to have pro-differentiation activity [64]. Nanog is also responsible for SMAD1 repression, thereby preventing BMP4-induced differentiation through the TGFbeta signaling pathway, for which SMAD1 is a key signal transducer [65].

Extraembryonic endoderm differentiation from ICM cells

Prior to embryo implantation one more differentiation takes place. Certain cells from the ICM give rise to the primitive endoderm, the first morphologically distinct cell type of the extraembryonic endoderm. The extraembryonic endoderm comprises the primitive, parietal and visceral endoderm components and will become the yolk sac during posterior development stages.

Wnt6 was recently identified as an inducer of primitive endoderm and this induction is accompanied by

translocation of beta-catenin (CTNNB1) and Snail1 to the nucleus [66]. This study also showed that up-regulation of protein kinase A (PKA) induces markers of parietal endoderm. Another Wnt family member, Wnt9a, is expressed only in ICM cells that surround the bastocoel [67] and induces repositioning of the cells expressing GATA6, which is necessary for formation of primitive endoderm [68].

Sox7 plays a major role in parietal endoderm differentiation. Through studies with short interfering RNA molecules, it was established that Sox7 is responsible for transcription induction of GATA4 and GATA6 [69]. Individual or combined silencing of Sox7, GATA4 and GATA6 result in suppression of cell shape changes and production of laminin-1 (LAMB1), characteristic changes present in parietal endoderm differentiation [69]. Gata4 was previously identified as a transcription factor responsible for the activation of FGF3 [70]. Sox7 also activates the FGF3 promoter. Conversely, Sox2 can negatively modulate the GATA4-dependent activation of FGF3, which is supported by the role of this factor in ICM pluripotency [71]. Another Sox family member, Sox17, is responsible for the differentiation of the extraembryonic endoderm in the final steps of preimplantation development [72]. The Runx1 factor is associated with the expression of Sox17 and is also specific for the extraembryonic endoderm [73]. HNF4 is a transcription factor specific of the extraembryonic endoderm with subsequent roles in post-implantation development and organogenesis [74]. Its expression may result from BMP4-induced differentiation [75]. Finally, the Dab2 protein is indispensable for the development of visceral endoderm; though its exact role is still not established, it is perhaps related to correct cell positioning [76,77]. The expression of Cer1, a marker of the anterior visceral endoderm (AVE), commences before embryo implantation in the subset of cells that comprise the primitive endoderm. This ancestral population includes both cells expressing Cer1 together with cells in which Cer1 expression begins after implantation and formation of the AVE [60].

Search for homologues

To establish an ortholog database and provide sequence information to the genes contained in the preimplantation pathway, aminoacid sequences corresponding to the human and mouse gene products were used as seed for the software SeedServer (Guedes *et al*., unpublished, see Methods for details). In fact, only the UniProt identifier for these proteins is necessary to execute SeedServer - gene symbols were verified in the NCBI Gene database and converted to the corresponding geneID, and the desired identifiers were obtained afterwards from the UniProt database. For each gene a cluster of homologues was generated comprising from 2 to 260 sequences (Additional file 1).

The recruited sequences contained in each cluster can be Swiss-Prot annotated or unrevised TrEMBL sequences. In total, 25% of the cluster sequences are Swiss-Prot, the great majority of clusters being comprised of TrEMBL sequences (75%). The search for homologues through SeedServer provides therefore a large amount of candidates for manual curation in Swiss-Prot. Furthermore, SeedServer can recruit sequences from organisms without a complete genome due to its use of UEKO (UEKO is built on top of Kegg Orthology homologues as UECOG [8] has been built on top of COG database) and bidirectional best hit (BBH) searches conducted by SeedLinkage [78], and in fact only 27% of the sequences present in all clusters are from organisms with a complete genome. The ortholog clustering by SeedServer was only performed for genes that had a corresponding SwissProt annotated gene product to be used as seed, therefore hRSCP and DPPA1, which are described in the pathway, did not go through this analysis.

Pathway ancestry

We then focused on the putative origin of these genes, determining which clade in the human lineage (e.g. class, order, family) shares each gene. The generation of ortholog clusters allowed for the determination of the last common ancestor (LCA) for each of the genes in the pathway. Figure 2 shows the genes according to their origin. Genes were arbitrarily considered ancient for this analysis if their last common ancestor originated before the divergence of the clade Euteleostomi and are coloured grey. Genes with a LCA belonging to the clade Euteleostomi or originated after divergence of Euteleostomi are considered recent genes and are coloured blue. Ancient origin genes with an ortholog in *Drosophila melanogaster* are marked with a red asterisk. This arbitrary classification was meant to attract attention to the two key pluripotency controlling genes, Nanog (ancient) and Oct4 (modern).

The graph shown in Figure 3 represents the distribution of all the genes in the pathway according to their origin respect to clades of the human lineage. It may be observed that a large quantity of genes originates in certain periods as seen in Eumetazoa, Coelomata, Euteleostomi and Eutheria. The reasons for this wavelike origin need to be further analysed. On the other hand, the apparent origin of complex structures, that characterize all descendents from a certain moment of evolution, might have occurred simultaneously to the specialization of gene groups. The coverage of genomic sequences in the database is far from homogeneous and can influence the shape of this graph [79]. In any case, the pattern

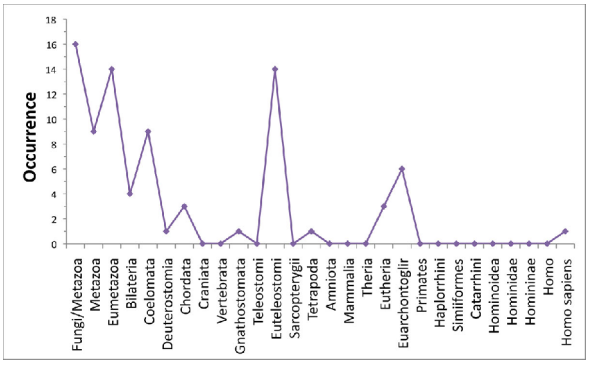


Figure 3 Gene origin in human evolution. Distribution of the genes in the preimplantation pathway according to their origin in clades of the human lineage, based on the determination of the Last Common Ancestor for the ortholog clusters generated by SeedServer. The y-axis represents the number of genes and the x-axis represents the taxonomical groups in which the genes originated.

observed agrees with the expansion of protein families related to stem cell markers observed in the ray-finned fish, that is, after divergence of the Euteleostomi [80].

Furthermore, we searched for functional information related to the *D. melanogaster* orthologues in order to determine if these functions are somehow similar or related to the functions of the corresponding pathway genes. This was done through a second text mining approach similar to the first and from the information recovered a secondary pathway was generated simply to illustrate the ortholog genes and their relative functional roles (Additional file 2). The regulatory pathways in which these genes are involved show us that these genes are all related to some part of *Drosophila* embryo development, some of them with highly conserved functions still observed in the preimplantation pathway described. An example is the Hippo signalling pathway, which is extremely conserved, showing Wts (Lats ortholog) phosphorylating Yki (Yap ortholog); this modification prevents Yki interaction with Sd (Tead4 ortholog). The correlation between the human gene names and corresponding *D. melanogaster* ortholog names can be found in Additional file 3 and also the PMID reference for the gene function in Drosophila development.

Discussion

The use of text-mining tools for the generation of regulatory pathways is an effective approach and it is important for the current interest of gathering data related to an organism or biological process. The search for information related to a specific concept such as "preimplantation development" resulted in the selection of data related to this process only. When other tools such as iHOP [1] and STRING [2] are used for the search of biointeractions, it is necessary to know the names for the genes you are interested on and the information is then retrieved. Moreover in the case of iHOP, the information retrieved consists of a large list of papers related to the gene of interest, which need to be manually analysed to extract the information related to the specific process. In the case of STRING, the result of a query is a network of direct associations to other genes, which can be activations, repressions, or unknown, but for which it is not possible to perform a search restricting the query to a specific process for which you seek to determine the involvement of a given gene.

The approach described in this work (using PubMed, MedlineRanker, PESCADOR) summarized in Figure 4, allows the researcher to initiate the study of a pathway without knowing exactly the genes involved, simply by selecting the published information related to the process of interest. The manual curation required to create a pathway through this approach is significantly smaller. However, the verification of all the interactions highlighted by the tool is essential. Text-mining is not able to eliminate the selection of false interaction pairs; in the case of LAITOR (contained in the PESCADOR platform), the type 3 and 4 interactions can present genes with no association specified in the text [20].

The text-mining data contribute the complete description of the pathway in the form of a literature review, a necessary step for the validation of the regulations represented, and for the inclusion of the pathway in a specific database, such as KEGG Pathway. The establishment of this procedure for pathway generation allows future work to enlarge the knowledge on subjects still not approached, such as regulatory pathways for several types of cancer, mechanisms of pathogen resistance in plants and response to abiotic stresses in plants, among other themes of interest.

The inclusion of the preimplantation pathway in databases such as the KEGG database will allow automatic annotation for several other organisms, as it is usually done in this database. Concurrently, a laboratory with a specific interest can promptly build a similar Pathway for its local use. From the 86 genes present in the pathway, 20 do not possess entries in KEGG Orthology and would constitute important additions. Considering that the contribution of KEGG for the sequence recruitment in the SeedServer clusters is only 25% of the total number of sequences, some organisms evolutionarily divergent from the ones represented in KEGG begin to play a more relevant role for a more efficient annotation of new sequences. It is relevant to stress that only the SeedLinkage and UEKO components of SeedServer are capable of clustering sequences proceeding from organisms without a complete genome project. Moreover, linkage

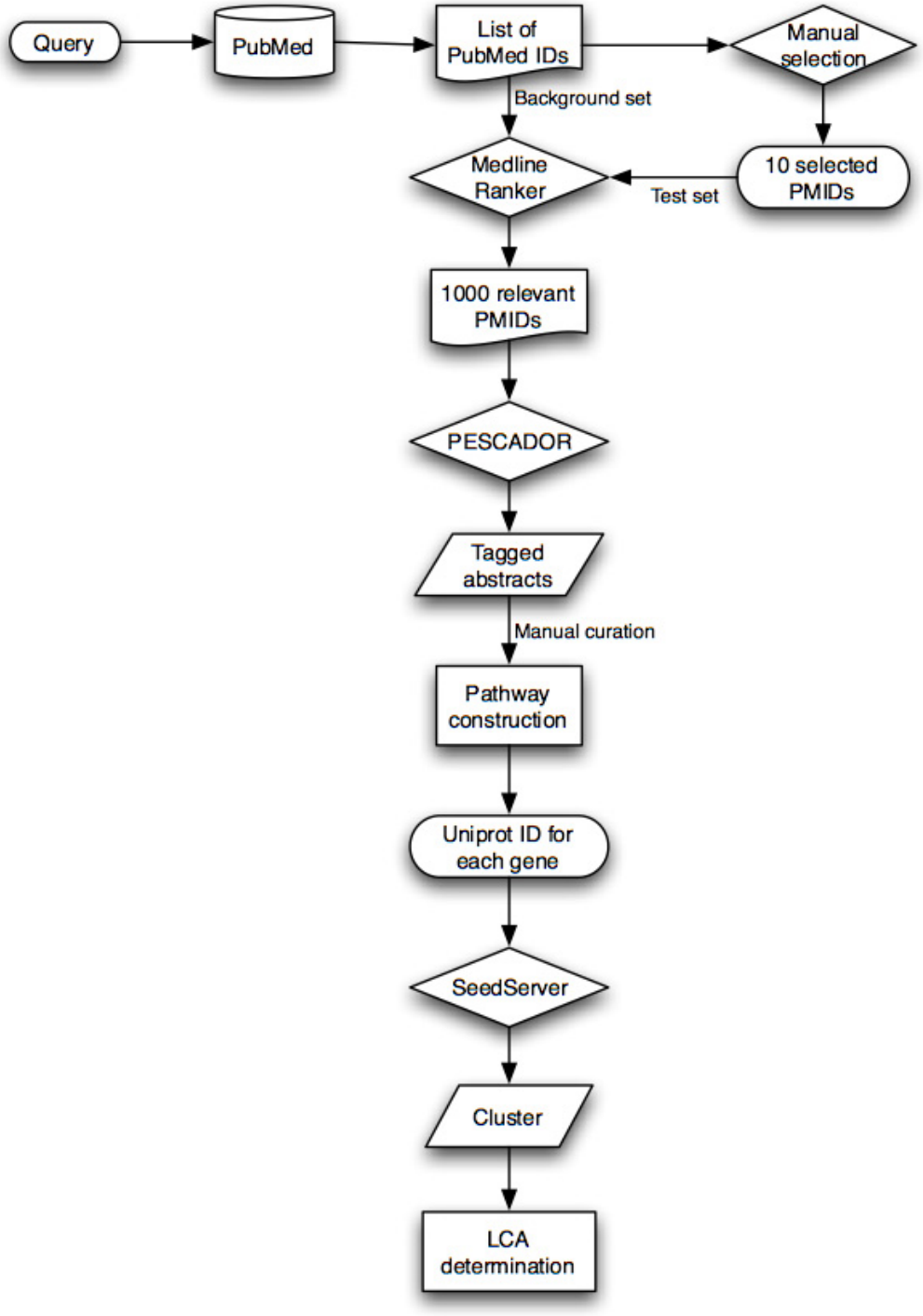


Figure 4 Pathway construction flowchart. The initial step consists of a PubMed search with the subject of interest (e.g. preimplantation development). The list of PubMed identifiers (PMIDs) obtained in the search is then used in the web tool Medline Ranker as the background set along with a list of PMIDs of manually selected abstracts considered informative which form the test set. The tool generates a list of abstracts classified by order of relevance. Best 1000 abstracts are recovered and their corresponding PMID is then introduced in the PESCADOR platform. Abstracts are tagged by PESCADOR and provide a source of biointeractions for manual curation and pathway construction. UniProt IDs for products of the genes present in the final pathway are obtained and used as seed in SeedServer. The software recruits homologues for each gene and creates the final clusters. Taxonomy IDs from each cluster can be used for Last Common Ancestor (LCA) determination.

of recruited to seed sequences are verified with PSI-BLAST.

Another important contribution from the ortholog clustering by SeedServer is the identification of candidates for Swiss-Prot Annotation. Swiss-Prot annotation depends on the correct association of sequences to gene families and proteins with known function, using the available literature as a reference. The annotation is facilitated since each of the genes is associated with PubMed Identifiers (PMIDs) stored in the PESCADOR tool, which are important references for the related orthologs.

The search for functional information for the *D. melanogaster* orthologues revealed the involvement of the genes in processes related to the embryonic development and was also a good validation for the clustering by SeedServer, since all sequences from *D. melanogaster* that clustered to the initial human and mouse genes present an embryo development related function.

Generation of correct clusters is essential for the correct determination of gene ancestry, but it is not the sole limiting factor. Sequencing of key organisms from taxonomic outgroups relative to the ones with complete genome sequences available will be a crucial source of sequences that will allow a revaluation of gene ancestry. Meanwhile, additional sequences clustered by software (SeedLinkage) and database enrichment (UEKO) improve the inspection of ancestry.

Determination of the ancestry for the genes in the preimplantation pathway was nonetheless a central analysis, given the expectancy that this pathway would be mainly formed by more contemporary components. Our data suggest that an ancient fraction of the pathway including Nanog and Sox2 originated before Chordata, whereas a modern fraction including Oct4 and LIF has appeared near the origin of Eutheria, the placentary organisms. Thus, an important transcriptional pathway comprising ancient and modern members has been characterized with text mining, and homologues search with SeedServer promptly allowed LCA determination.

Conclusions

Generation of regulatory pathways through text-mining tools allows integration of data generated by previous studies for a more complete view of a biological process. If the genes present in this pathway are associated with clusters of orthologues this information is added to the pathway making the visualization of the same process available for different organisms. The analysis of orthology also permits determination of the ancestry of the genes involved in the process leading to a better understanding of the evolution of such process.

Methods

Text-mining and pathway construction

NCBI's PubMed database was used as a source of available literature (http://www.ncbi.nlm.nih.gov/pubmed) for the text-mining approach. The search query used was "preimplantation development" and the PubMed identification numbers of the selected papers (PMIDs) were saved as a text file. Ten papers were selected manually by us to be used in the Medline Ranker software ([19]; http://cbdm.mdc-berlin.de/tools/medlineranker/). These papers, (references [23,26,28,29,31,50, 59,81,82]), were considered by us as highly informative because they described numerous gene regulations concerning preimplantation development. We used the PMIDs retrieved by the PubMed search as the background set and the 10 manually selected PMIDs as the training set. After classification by order of relevance we selected the 1000 better-classified abstracts for further analysis presenting a p-value < 0.01. These abstracts were then submitted through PESCADOR (manuscript under preparation, Barbosa-Silva *et al.*), an online platform for the software LAITOR [20]. After PESCADOR, results were manually curated and the gene biointeractions recovered were used to build a regulatory pathway in Keynote MacOS according to the markup language used by KEGG for pathway construction (KGML can be found at http://www.genome.jp/kegg/xml/docs/). This process consisted mainly of finding the highlighted interaction in the abstract tagged by PESCADOR, confirming its involvement in the preimplantation development by checking the corresponding paper and drawing this interaction in the pathway picture.

SeedServer search for homologues

UniProt IDs for human and mouse gene products corresponding to each of the genes represented in the preimplantation pathway were used as seed in the SeedServer software (not published, Guedes et al.). SeedServer is a web application (http://biodados.icb.ufmg.br/seedserver/) which searches for homologous sequences through two components: the program SeedLinkage [78] and the databases KEGG Orthology (KO) [5] and its enriched version UEKO (unpublished, developed by Fernandes *et al.* by application of the procedure described to enrich COG [10] to the KEGG Orthology database). Clustering was verified by PSI-BLAST searches using seed sequences as query and the recruited proteins as database, and eventual false positives were discarded (1.5% of the recruited sequences).

LCA determination

Clusters generated for each of the pathway genes were used to determine the Last Common Ancestor (LCA) of

each gene. Each cluster provided a list of Taxonomy IDs corresponding to the organisms in which orthologs of the pathway genes were found. The clade in the human lineage that comprised these Taxonomy IDs as leaves in the Taxonomy Tree was considered to bear the LCA.

Note added in proof

PESCADOR, referred in the text as in preparation, is now published: PESCADOR, a web-based tool to assist text-mining of biointeractions extracted from PubMed queries. Barbosa-Silva A, Fontaine JF, Donnard ER, Stussi F, Ortega JM, Andrade-Navarro MA. BMC Bioinformatics. 2011 Nov 9;12(1):435. [Epub ahead of print] PMID: 22070195[83].

Additional material

Additional file 1: Homolog clusters. Clusters of homologous sequences found by SeedServer for each of the genes in the preimplantation pathway. For each gene, the left column shows the clustered sequence Uniprot ID and the right column shows the Taxonomy ID for this sequence.

Additional file 2: Ortholog functions in *Drosophila melanogaster*. This figure represents the corresponding *D. melanogaster* orthologs found by SeedServer and their respective interactions and functions in fruit fly development. Note that these orthologs are involved in processes related to *D. melanogaster* embryo development. See Additional file 3 for a table with gene name correspondence between the genes in this figure and the ones on Figure 3.

Additional file 3: Gene correspondence table. Human and *Drosophila melanogaster* gene name correspondence for the orthologs grouped by SeedServer. Column 3 lists the PubMed identifiers (PMIDs) from the papers where functions described in Additional file 2 were found.

Acknowledgements

This article has been published as part of *BMC Genomics* Volume 12 Supplement 4, 2011: Proceedings of the 6th International Conference of the Brazilian Association for Bioinformatics and Computational Biology (X-meeting 2010). The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2164/12?issue=S4

Author details

[1]Laboratório Biodados, Dept. de Bioquímica e Imunologia, Universidade Federal de Minas Gerais, 31270-901, Belo Horizonte - MG, Brazil. [2]Departamento de Bioquímica, Universidade de São Paulo - SP, Brazil. [3]Computational Biology and Data Mining Group, Max-Delbrück Center for Molecular Medicine, Robert-Rössle-Strasse 10, D-13125, Berlin, Germany. [4]Bioinformatics Graduate Program, Federal University of Paraná - UFPR (SEPT). Rua Dr. Alcides Vieira Arcoverde 1225, CEP 81520-260. Curitiba-PR, Brazil. [5]New York State Stem Cell Science, New York State Department of Health Wadsworth Center, Rm C345, New York, USA.

Authors' contributions

ERD and JMO conceived the project and wrote the paper. ERD performed the research and pathway construction. MJK curated the pathway biointeractions. ABS (author of SeedLinkage and LAITOR) and MAAN designed the PESCADOR platform. RLMG designed the SeedServer software and conducted the ortholog search. HV was responsible for the LCA determination. GRF constructed the UEKO database. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Published: 22 December 2011

References

1. Hoffmann R, Valencia A: **A gene network for navigating the literature.** *Nat Genet* 2004, **36**:664.
2. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, *et al*: **STRING 8–a global view on proteins and their functional interactions in 630 organisms.** *Nucleic Acids Res* 2009, **37**:D412-416.
3. Letunic I, Yamada T, Kanehisa M, Bork P: **iPath: interactive exploration of biochemical pathways and networks.** *Trends Biochem Sci* 2008, **33**:101-103.
4. Karp PD, Ouzounis CA, Moore-Kochlacs C, Goldovsky L, Kaipa P, Ahren D, Tsoka S, Darzentas N, Kunin V, Lopez-Bigas N: **Expansion of the BioCyc collection of pathway/genome databases to 160 genomes.** *Nucleic Acids Res* 2005, **33**:6083-6089.
5. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res* 2000, **28**:27-30.
6. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, *et al*: **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics* 2003, **4**:41.
7. Muller J, Szklarczyk D, Julien P, Letunic I, Roth A, Kuhn M, Powell S, von Mering C, Doerks T, Jensen LJ, Bork P: **eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations.** *Nucleic Acids Res* 2010, **38**:D190-195.
8. Li L, Stoeckert CJ Jr., Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Res* 2003, **13**:2178-2189.
9. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH: **UniRef: comprehensive and non-redundant UniProt reference clusters.** *Bioinformatics* 2007, **23**:1282-1288.
10. Fernandes GR, Barbosa DV, Prosdocimi F, Pena IA, Santana-Santos L, Coelho Junior O, Barbosa-Silva A, Velloso HM, Mudado MA, Natale DA, *et al*: **A procedure to recruit members to enlarge protein family databases–the building of UECOG (UniRef-Enriched COG Database) as a model.** *Genet Mol Res* 2008, **7**:910-924.
11. Guo G, Huss M, Tong GQ, Wang C, Li Sun L, Clarke ND, Robson P: **Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst.** *Dev Cell* 2010, **18**:675-685.
12. Medvedev SP, Shevchenko AI, Mazurok NA, Zakiian SM: **[OCT4 and NANOG are the key genes in the system of pluripotency maintenance in mammalian cells].** *Genetika* 2008, **44**:1589-1608.
13. Tam WL, Lim CY, Han J, Zhang J, Ang YS, Ng HH, Yang H, Lim B: **T-cell factor 3 regulates embryonic stem cell pluripotency and self-renewal by the transcriptional control of multiple lineage pathways.** *Stem Cells* 2008, **26**:2019-2031.
14. Yamanaka Y, Ralston A, Stephenson RO, Rossant J: **Cell and molecular regulation of the mouse blastocyst.** *Dev Dyn* 2006, **235**:2301-2314.
15. Zhou Q, Chipperfield H, Melton DA, Wong WH: **A gene regulatory network in mouse embryonic stem cells.** *Proc Natl Acad Sci USA* 2007, **104**:16438-16443.
16. Wang H, Dey SK: **Roadmap to embryo implantation: clues from mouse models.** *Nat Rev Genet* 2006, **7**:185-199.
17. Johnson MH, McConnell JM: **Lineage allocation and cell polarity during mouse embryogenesis.** *Semin Cell Dev Biol* 2004, **15**:583-597.
18. Davidson EH, Rast JP, Oliveri P, Ransick A, Calestani C, Yuh CH, Minokawa T, Amore G, Hinman V, Arenas-Mena C, *et al*: **A genomic regulatory network for development.** *Science* 2002, **295**:1669-1678.
19. Fontaine JF, Barbosa-Silva A, Schaefer M, Huska MR, Muro EM, Andrade-Navarro MA: **MedlineRanker: flexible ranking of biomedical literature.** *Nucleic Acids Res* 2009, **37**:W141-146.
20. Barbosa-Silva A, Soldatos TG, Magalhaes IL, Pavlopoulos GA, Fontaine JF, Andrade-Navarro MA, Schneider R, Ortega JM: **LAITOR–Literature Assistant for Identification of Terms co-Occurrences and Relationships.** *BMC Bioinformatics* 2010, **11**:70.

21. Suzuki T, Abe K, Inoue A, Aoki F: **Expression of c-MYC in nuclear speckles during mouse oocyte growth and preimplantation development.** *J Reprod Dev* 2009, **55**:491-495.
22. Monk M, Hitchins M, Hawes S: **Differential expression of the embryo/cancer gene ECSA(DPPA2), the cancer/testis gene BORIS and the pluripotency structural gene OCT4, in human preimplantation development.** *Molecular Human Reproduction* 2008, **14**:347-355.
23. Mizuno S, Sono Y, Matsuoka T, Matsumoto K, Saeki K, Hosoi Y, Fukuda A, Morimoto Y, Iritani A: **Expression and subcellular localization of GSE protein in germ cells and preimplantation embryos.** *J Reprod Dev* 2006, **52**:429-438.
24. Chung YG, Ratnam S, Chaillet JR, Latham KE: **Abnormal regulation of DNA methyltransferase expression in cloned mouse embryos.** *Biol Reprod* 2003, **69**:146-153.
25. Yu JN, Xue CY, Wang XG, Lin F, Liu CY, Lu FZ, Liu HL: **5-AZA-2'-deoxycytidine (5-AZA-CdR) leads to down-regulation of Dnmt1o and gene expression in preimplantation mouse embryos.** *Zygote* 2009, **17**:137-145.
26. Plusa B, Frankenberg S, Chalmers A, Hadjantonakis AK, Moore CA, Papalopulu N, Papaioannou VE, Glover DM, Zernicka-Goetz M: **Downregulation of Par3 and aPKC function directs cells towards the ICM in the preimplantation mouse embryo.** *J Cell Sci* 2005, **118**:505-515.
27. Ralston A, Rossant J: **Cdx2 acts downstream of cell polarization to cell-autonomously promote trophectoderm fate in the early mouse embryo.** *Dev Biol* 2008, **313**:614-629.
28. Nishioka N, Inoue K, Adachi K, Kiyonari H, Ota M, Ralston A, Yabuta N, Hirahara S, Stephenson RO, Ogonuki N, *et al*: **The Hippo signaling pathway components Lats and Yap pattern Tead4 activity to distinguish mouse trophectoderm from inner cell mass.** *Dev Cell* 2009, **16**:398-410.
29. Strumpf D, Mao CA, Yamanaka Y, Ralston A, Chawengsaksophak K, Beck F, Rossant J: **Cdx2 is required for correct cell fate specification and differentiation of trophectoderm in the mouse blastocyst.** *Development* 2005, **132**:2093-2102.
30. Hyslop L, Stojkovic M, Armstrong L, Walter T, Stojkovic P, Przyborski S, Herbert M, Murdoch A, Strachan T, Lako M: **Downregulation of NANOG induces differentiation of human embryonic stem cells to extraembryonic lineages.** *Stem Cells* 2005, **23**:1035-1043.
31. Zhang J, Tam WL, Tong GQ, Wu Q, Chan HY, Soh BS, Lou Y, Yang J, Ma Y, Chai L, *et al*: **Sall4 modulates embryonic stem cell pluripotency and early embryonic development by the transcriptional regulation of Pou5f1.** *Nat Cell Biol* 2006, **8**:1114-1123.
32. Ralston A, Cox BJ, Nishioka N, Sasaki H, Chea E, Rugg-Gunn P, Guo G, Robson P, Draper JS, Rossant J: **Gata3 regulates trophoblast development downstream of Tead4 and in parallel to Cdx2.** *Development* 2010, **137**:395-403.
33. Vauti F, Prochnow BR, Freese E, Ramasamy SK, Ruiz P, Arnold HH: **Arp3 is required during preimplantation development of the mouse embryo.** *FEBS Lett* 2007, **581**:5691-5697.
34. Hayashi Y, Furue MK, Tanaka S, Hirose M, Wakisaka N, Danno H, Ohnuma K, Oeda S, Aihara Y, Shiota K, *et al*: **BMP4 induction of trophoblast from mouse embryonic stem cells in defined culture conditions on laminin.** *In Vitro Cell Dev Biol Anim* 2010, **46**:416-430.
35. Kondo M, Cubillo E, Tobiume K, Shirakihara T, Fukuda N, Suzuki H, Shimizu K, Takehara K, Cano A, Saitoh M, Miyazono K: **A role for Id in the regulation of TGF-beta-induced epithelial-mesenchymal transdifferentiation.** *Cell Death Differ* 2004, **11**:1092-1101.
36. Riley P, Anson-Cartwright L, Cross JC: **The Hand1 bHLH transcription factor is essential for placentation and cardiac morphogenesis.** *Nat Genet* 1998, **18**:271-275.
37. Babaie Y, Herwig R, Greber B, Brink TC, Wruck W, Groth D, Lehrach H, Burdon T, Adjaye J: **Analysis of Oct4-dependent transcriptional networks regulating self-renewal and pluripotency in human embryonic stem cells.** *Stem Cells* 2007, **25**:500-510.
38. Marikawa Y, Alarcón VB: **Establishment of trophectoderm and inner cell mass lineages in the mouse embryo.** *Mol Reprod Dev* 2009, **76**:1019-1032.
39. Adjaye J, Herwig R, Brink TC, Herrmann D, Greber B, Sudheer S, Groth D, Carnwath JW, Lehrach H, Niemann H: **Conserved molecular portraits of bovine and human blastocysts as a consequence of the transition from maternal to embryonic control of gene expression.** *Physiol Genomics* 2007, **31**:315-327.
40. Shin MR, Cui XS, Jun JH, Jeong YJ, Kim NH: **Identification of mouse blastocyst genes that are downregulated by double-stranded RNA-mediated knockdown of Oct-4 expression.** *Mol Reprod Dev* 2005, **70**:390-396.
41. Yamada K, Ogawa H, Tamiya G, Ikeno M, Morita M, Asakawa S, Shimizu N, Okazaki T: **Genomic organization, chromosomal localization, and the complete 22 kb DNA sequence of the human GCMa/GCM1, a placenta-specific transcription factor gene.** *Biochem Biophys Res Commun* 2000, **278**:134-139.
42. Matin MM, Walsh JR, Gokhale PJ, Draper JS, Bahrami AR, Morton I, Moore HD, Andrews PW: **Specific knockdown of Oct4 and beta2-microglobulin expression by RNA interference in human embryonic stem cells and embryonic carcinoma cells.** *Stem Cells* 2004, **22**:659-668.
43. Collins JE, Lorimer JE, Garrod DR, Pidsley SC, Buxton RS, Fleming TP: **Regulation of desmocollin transcription in mouse preimplantation embryos.** *Development* 1995, **121**:743-753.
44. Chen L, Yang M, Dawes J, Khillan JS: **Suppression of ES cell differentiation by retinol (vitamin A) via the overexpression of Nanog.** *Differentiation* 2007, **75**:682-693.
45. Jiang J, Chan YS, Loh YH, Cai J, Tong GQ, Lim CA, Robson P, Zhong S, Ng HH: **A core Klf circuitry regulates self-renewal of embryonic stem cells.** *Nat Cell Biol* 2008, **10**:353-360.
46. Pan G, Li J, Zhou Y, Zheng H, Pei D: **A negative feedback loop of transcription factors that controls stem cell pluripotency and self-renewal.** *FASEB J* 2006, **20**:1730-1732.
47. van den Berg DL, Zhang W, Yates A, Engelen E, Takacs K, Bezstarosti K, Demmers J, Chambers I, Poot RA: **Estrogen-related receptor beta interacts with Oct4 to positively regulate Nanog gene expression.** *Mol Cell Biol* 2008, **28**:5986-5995.
48. Zhang X, Zhang J, Wang T, Esteban MA, Pei D: **Esrrb activates Oct4 transcription and sustains self-renewal and pluripotency in embryonic stem cells.** *J Biol Chem* 2008, **283**:35825-35833.
49. Okumura-Nakanishi S, Saito M, Niwa H, Ishikawa F: **Oct-3/4 and Sox2 regulate Oct-3/4 gene in embryonic stem cells.** *J Biol Chem* 2005, **280**:5307-5317.
50. Cauffman G, De Rycke M, Sermon K, Liebaers I, Van de Velde H: **Markers that define stemness in ESC are unable to identify the totipotent cells in human preimplantation embryos.** *Hum Reprod* 2009, **24**:63-70.
51. Holzinger M, Bouffier L, Villalonga R, Cosnier S: **Adamantane/beta-cyclodextrin affinity biosensors based on single-walled carbon nanotubes.** *Biosens Bioelectron* 2009, **24**:1128-1134.
52. Adjaye J, Huntriss J, Herwig R, BenKahla A, Brink TC, Wierling C, Hultschig C, Groth D, Yaspo ML, Picton HM, *et al*: **Primary differentiation in the human blastocyst: comparative molecular portraits of inner cell mass and trophectoderm cells.** *Stem Cells* 2005, **23**:1514-1525.
53. Cormier S, Le Bras S, Souilhol C, Vandormael-Pournin S, Durand B, Babinet C, Baldacci P, Cohen-Tannoudji M: **The murine ortholog of notchless, a direct regulator of the notch pathway in Drosophila melanogaster, is essential for survival of inner cell mass cells.** *Mol Cell Biol* 2006, **26**:3541-3549.
54. Voss AK, Thomas T, Petrou P, Anastassiadis K, Schöler H, Gruss P: **Taube nuss is a novel gene essential for the survival of pluripotent cells of early mouse embryos.** *Development* 2000, **127**:5449-5461.
55. Robson P, Stein P, Zhou B, Schultz RM, Baldwin HS: **Inner cell mass-specific expression of a cell adhesion molecule (PECAM-1/CD31) in the mouse blastocyst.** *Dev Biol* 2001, **234**:317-329.
56. Sun C, Nakatake Y, Akagi T, Ura H, Matsuda T, Nishiyama A, Koide H, Ko MS, Niwa H, Yokota T: **Dax1 binds to Oct3/4 and inhibits its transcriptional activity in embryonic stem cells.** *Mol Cell Biol* 2009, **29**:4574-4583.
57. Pereira L, Yi F, Merrill BJ: **Repression of Nanog gene transcription by Tcf3 limits embryonic stem cell self-renewal.** *Mol Cell Biol* 2006, **26**:7479-7491.
58. Tanaka TS, Lopez de Silanes I, Sharova LV, Akutsu H, Yoshikawa T, Amano H, Yamanaka S, Gorospe M, Ko MS: **Esg1, expressed exclusively in preimplantation embryos, germline, and embryonic stem cells, is a putative RNA-binding protein with broad RNA targets.** *Dev Growth Differ* 2006, **48**:381-390.
59. Ambrosetti DC, Schöler HR, Dailey L, Basilico C: **Modulation of the activity of multiple transcriptional activation domains by the DNA binding domains mediates the synergistic action of Sox2 and Oct-3 on the fibroblast growth factor-4 enhancer.** *J Biol Chem* 2000, **275**:23387-23397.

60. Torres-Padilla ME, Richardson L, Kolasinska P, Meilhac SM, Luetke-Eversloh MV, Zernicka-Goetz M: **The anterior visceral endoderm of the mouse embryo is established from both preimplantation precursor cells and by de novo gene expression after implantation.** *Dev Biol* 2007, **309**:97-112.
61. Chakravarthy H, Boer B, Desler M, Mallanna SK, McKeithan TW, Rizzino A: **Identification of DPPA4 and other genes as putative Sox2:Oct-3/4 target genes using a combination of in silico analysis and transcription-based assays.** *J Cell Physiol* 2008, **216**:651-662.
62. Saito S, Liu B, Yokoyama K: **Animal embryonic stem (ES) cells: self-renewal, pluripotency, transgenesis and nuclear transfer.** *Hum Cell* 2004, **17**:107-115.
63. De Felici M, Farini D, Dolci S: **In or out stemness: comparing growth factor signalling in mouse embryonic stem cells and primordial germ cells.** *Curr Stem Cell Res Ther* 2009, **4**:87-97.
64. Torres J, Watt FM: **Nanog maintains pluripotency of mouse embryonic stem cells by inhibiting NFkappaB and cooperating with Stat3.** *Nat Cell Biol* 2008, **10**:194-201.
65. Suzuki A, Raya A, Kawakami Y, Morita M, Matsui T, Nakashima K, Gage FH, Rodríguez-Esteban C, Izpisúa Belmonte JC: **Nanog binds to Smad1 and blocks bone morphogenetic protein-induced differentiation of embryonic stem cells.** *Proc Natl Acad Sci USA* 2006, **103**:10294-10299.
66. Krawetz R, Kelly GM: **Wnt6 induces the specification and epithelialization of F9 embryonal carcinoma cells to primitive endoderm.** *Cell Signal* 2008, **20**:506-517.
67. Kemp C, Willems E, Abdo S, Lambiv L, Leyns L: **Expression of all Wnt genes and their secreted antagonists during mouse blastocyst and postimplantation development.** *Dev Dyn* 2005, **233**:1064-1075.
68. Meilhac SM, Adams RJ, Morris SA, Danckaert A, Le Garrec JF, Zernicka-Goetz M: **Active cell movements coupled to positional induction are involved in lineage segregation in the mouse blastocyst.** *Dev Biol* 2009, **331**:210-221.
69. Futaki S, Hayashi Y, Emoto T, Weber CN, Sekiguchi K: **Sox7 plays crucial roles in parietal endoderm differentiation in F9 embryonal carcinoma cells through regulating Gata-4 and Gata-6 expression.** *Mol Cell Biol* 2004, **24**:10492-10503.
70. Murakami A, Thurlow J, Dickson C: **Retinoic acid-regulated expression of fibroblast growth factor 3 requires the interaction between a novel transcription factor and GATA-4.** *J Biol Chem* 1999, **274**:17242-17248.
71. Murakami A, Shen H, Ishida S, Dickson C: **SOX7 and GATA-4 are competitive activators of Fgf-3 transcription.** *J Biol Chem* 2004, **279**:28564-28573.
72. Shimoda M, Kanai-Azuma M, Hara K, Miyazaki S, Kanai Y, Monden M, Miyazaki J: **Sox17 plays a substantial role in late-stage differentiation of the extraembryonic endoderm in vitro.** *J Cell Sci* 2007, **120**:3859-3869.
73. Kurimoto K, Yabuta Y, Ohinata Y, Ono Y, Uno KD, Yamada RG, Ueda HR, Saitou M: **An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis.** *Nucleic Acids Res* 2006, **34**:e42.
74. Duncan SA, Manova K, Chen WS, Hoodless P, Weinstein DC, Bachvarova RF, Darnell JE: **Expression of transcription factor HNF-4 in the extraembryonic endoderm, gut, and nephrogenic tissue of the developing mouse embryo: HNF-4 is a marker for primary endoderm in the implanting blastocyst.** *Proc Natl Acad Sci USA* 1994, **91**:7598-7602.
75. Coucouvanis E, Martin GR: **BMP signaling plays a role in visceral endoderm differentiation and cavitation in the early mouse embryo.** *Development* 1999, **126**:535-546.
76. Morris SM, Tallquist MD, Rock CO, Cooper JA: **Dual roles for the Dab2 adaptor protein in embryonic development and kidney transport.** *EMBO J* 2002, **21**:1555-1564.
77. Yang DH, Smith ER, Roland IH, Sheng Z, He J, Martin WD, Hamilton TC, Lambeth JD, Xu XX: **Disabled-2 is essential for endodermal cell positioning and structure formation during mouse embryogenesis.** *Dev Biol* 2002, **251**:27-44.
78. Barbosa-Silva A, Satagopam VP, Schneider R, Ortega JM: **Clustering of cognate proteins among distinct proteomes derived from multiple links to a single seed sequence.** *BMC Bioinformatics* 2008, **9**:141.
79. Perez-Iratxeta C, Palidwor G, Andrade-Navarro MA: **Towards completion of the Earth's proteome.** *EMBO Rep* 2007, **8**:1135-1141.
80. Krzyzanowski PM, Andrade-Navarro MA: **Identification of novel stem cell markers using gap analysis of gene expression data.** *Genome Biol* 2007, **8**:R193.
81. Scaffidi P, Bianchi ME: **Spatially precise DNA bending is an essential activity of the sox2 transcription factor.** *J Biol Chem* 2001, **276**:47296-47302.
82. Lim CY, Tam WL, Zhang J, Ang HS, Jia H, Lipovich L, Ng HH, Wei CL, Sung WK, Robson P, *et al*: **Sall4 regulates distinct transcription circuitries in different blastocyst-derived stem cell lineages.** *Cell Stem Cell* 2008, **3**:543-554.
83. Barbosa-Silva A, Fontaine JF, Donnard ER, Stussi F, Ortega JM, Andrade-Navarro MA: **PESCADOR, a web-based tool to assist text-mining of biointeractions extracted from PubMed queries.** *BMC Bioinformatics* 2011, **12**:435.

doi:10.1186/1471-2164-12-S4-S3
Cite this article as: Donnard *et al.*: Preimplantation development regulatory pathway construction through a text-mining approach. *BMC Genomics* 2011 **12**(Suppl 4):S3.