MINISTÉRIO DA SAÚDE
FUNDAÇÃO OSWALDO CRUZ
INSTITUTO OSWALDO CRUZ

Doutorado em Programa de Pós-Graduação Biologia Computacional e Sistemas

ANALOGIA FUNCIONAL NO METABOLISMO HUMANO: ENZIMAS COM DISTINTOS PAPÉIS BIOLÓGICOS OU REDUNDÂNCIA FUNCIONAL?

RAFAEL MINA PIERGIORGE

Rio de Janeiro
Dezembro de 2017

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Programa de Pós-Graduação em Biologia Computacional e Sistemas

RAFAEL MINA PIERGIORGE

Analogia Funcional No Metabolismo Humano: Enzimas Com Distintos Papéis Biológicos Ou Redundância Funcional?

Tese apresentada ao Instituto Oswaldo Cruz como parte dos requisitos para obtenção do título de Doutor em Biologia Computacional e Sistemas

Orientador (es): Prof. Dr.Marcos Paulo Catanho de Souza
Prof. Dra. Ana Carolina Ramos Guimarães

RIO DE JANEIRO
Dezembro de 2017

Piergiorge, Rafael.

ANALOGIA FUNCIONAL NO METABOLISMO HUMANO: ENZIMAS COM DISTINTOS PAPÉIS BIOLÓGICOS OU REDUNDÂNCIA FUNCIONAL? / Rafael Piergiorge. - Rio de janeiro, 2017.
136 f.; il.

Tese (Doutorado) – Instituto Oswaldo Cruz, Pós-Graduação em Biologia Computacional e Sistemas, 2017.

Orientador: Marcos Catanho.
Co-orientadora: Ana Carolina Ramos Guimarães.

Bibliografia: f. 72-81

1. atividade enzimática. 2. evolução convergente. 3. mapeamento genômico. 4. alinhamento estrutural. 5. bioinformática. I. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da Biblioteca de Manguinhos/ICICT com os dados fornecidos pelo(a) autor(a).

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Programa de Pós-Graduação em Biologia Computacional e Sistemas

AUTOR: RAFAEL MINA PIERGIORGE

ANALOGIA FUNCIONAL NO METABOLISMO HUMANO: ENZIMAS COM DISTINTOS PAPÉIS BIOLÓGICOS OU REDUNDÂNCIA FUNCIONAL?

ORIENTADOR (ES): Prof. Dr. Marcos Paulo Catanho de Souza
Prof. Dr. Ana Carolina Ramos Guimarães

Aprovada em: _____/_____/_____

EXAMINADORES:

Prof. Dr. Rafael Dias Mesquita – Presidente (Instituto de Química/UFRJ)
Prof. Dr. Diogo AntonioTschoeke(UFRJ)
Prof. Dr. Thiago Estevam Parente Martins (IOC/FIOCRUZ)
Prof. Dr. Fabio Faria da Mota (IOC/FIOCRUZ)
Prof. Dra. Adriana Machado Fróes (INPI)

Rio de Janeiro, 14 de dezembro de 2017

Anexar a cópia da Ata que será entregue pela SEAC já assinada.

Em memória de Maria Helena Pereira
Mina Piergiorge

AGRADECIMENTOS

Aos meus pais, Luiz Piergiorge e Maria Helena Mina, por me ajudarem a concluir mais uma etapa da minha vida acadêmica.

Aos meus orientadores Marcos Catanho e Ana Carolina Guimarães, pelo voto de confiança que depositaram em mim. Vocês foram verdadeiros orientadores, principalmente nos momentos de dificuldades. Agradeço a ambos por tudo que aprendi ao longo desses quatro anos e o que espero continuar aprendendo, por meio de colaborações e amizade. Por me ajudarem a desenvolver meu senso crítico e ampliar minha curiosidade científica.

Aos amigos Alexander Franca, Mayla Abrahim, Edson Silva Machado, André Luiz Quintanilha Torres, July Linhares e Luiz Phillippe Ribeiro Baptista.

Aos amigos e pesquisadores do Laboratório de Genômica Funcional e Bioinformática – LAGFB.

Aos doutores Fabio Passetti e Natasha Jorge, pelo auxilio durante as análises de expressão gênica.

Ao programa CAPES-UDELAR, pela oportunidade de trabalhar durante cinco meses sob a supervisão do pesquisador Fernando Alvarez-Valín na Universidade da República - Uruguai.

À Plataforma de Bioinformática da Fiocruz RPT04A/RJ.

Ao Programa de Pós-Graduação em Biologia Computacional e Sistemas (BCS) e ao IOC por me possibilitarem uma formação sólida em bioinformática.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES e a Fundação Oswaldo Cruz pelo auxílio financeiro.

"Se vi mais longe foi por estar de pé sobre
ombros de gigantes."

Isaac Newton

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

ANALOGIA FUNCIONAL NO METABOLISMO HUMANO: ENZIMAS COM DISTINTOS PAPÉIS BIOLÓGICOS OU REDUNDÂNCIA FUNCIONAL?

RESUMO

TESE DE DOUTORADO EM BIOLOGIA COMPUTACIONAL E SISTEMAS

Rafael Mina Piergiorge

Uma vez que as enzimas catalisam quase todas as reações químicas que ocorrem nos organismos vivos, é crucial que os genes que codificam essas atividades sejam corretamente identificados e funcionalmente caracterizados. Estudos sugerem que a fração de atividades enzimáticas em que múltiplos eventos de origem independente ocorreram durante a evolução é substancial. Questões como qual a origem das enzimas análogas, por que tantos eventos de origem independente aparentemente ocorreram durante a evolução, e quais são os motivos da co-ocorrência no mesmo organismo de formas enzimáticas distintas que catalisam a mesma reação, permanecem sem resposta. Além disso, várias enzimas isofuncionais ainda não são reconhecidas como não-homólogas, mesmo com evidências indicando diferentes histórias evolutivas. Neste trabalho, propusemos investigar o papel biológico e evolutivo da existência de análogos funcionais intragenômicos em vias metabólicas na espécie humana. Foram encontradas evidências de enzimas isofuncionais não-homólogas catalisando reações anotadas em 15 atividades enzimáticas. Essas atividades enzimáticas estão associadas com nove vias/processos biológicos. Dessa forma, levantamos a hipótese de que 70 genes codificadores de enzimas análogas não deveria ser interpretada como redundância funcional, uma vez que estas enzimas análogas intragenômicas poderiam estar envolvidas em papéis biológicos distintos. Para testar esta hipótese, nós comparamos o perfil de transcrição desses genes que codificam o repertório de enzimas análogas intragenômicas catalogadas no metabolismo humano, utilizando dados RNA-Seq obtidos a partir de 8.555 amostras de 53 diferentes tecidos humanos saudáveis, publicamente disponíveis, além de identificar miRNAs que possivelmente estivessem envolvidos na modulação da expressão dos genes codificadores de enzimas análogas estudados, bem como reunimos informações sobre as localizações subcelulares destas enzimas. Os resultados das análises comparativas dos perfis de expressão, redes de interação com miRNAs e localização subcelular parecem refutar a hipótese sobre redundância funcional, já que existem distintos reguladores, localizações subcelulares e alternância no padrão de expressão dos genes codificadores de enzimas análogas catalisando uma determinada reação.

INSTITUTO OSWALDO CRUZ

FUNCTIONAL ANALOGY IN HUMAN METABOLISM: ENZYMES WITH DIFFERENT BIOLOGICAL ROLES OR FUNCTIONAL REDUNDANCY?

ABSTRACT

PHD THESIS IN COMPUTATIONAL AND SYSTEMS BIOLOGY

Rafael Mina Piergiorge

Since enzymes catalyze almost all chemical reactions that occur in living organisms, it is crucial that the genes coding for these activities be correctly identified and functionally characterized. Studies suggest that the fraction of enzymatic activities in which multiple events of independent origin occurred during evolution is substantial. Concerns such as the origin of analogous enzymes, why so many events of independent origin apparently occurred during evolution, and what are the reasons for the co-occurrence in the same organism of distinct enzymatic forms that catalyze the same reaction remain unanswered. In addition, several isofunctional enzymes are not yet recognized as non-homologous, even with evidence indicating different evolutionary histories. In this work, we propose to investigate the biological and evolutionary role of intragenomic functional analogues in metabolic pathways in human species. We found evidence of non-homologous isofunctional enzymes catalyzing reactions noted in 15 enzymatic activities. These enzymatic activities are associated with nine pathway/biological processes. Thus, we hypothesized that 70 genes encoding analogous enzymes should not be interpreted as functional redundancy since these analogous intragenomic enzymes could be involved in different biological roles. To test this hypothesis, we compared the transcription profile of these genes encoding the repertoire of analogous intragenomic enzymes cataloged in human metabolism using RNA-Seq data obtained from 8,555 samples from 53 different publicly available healthy human tissues, in addition to identifying miRNAs that might be involved in modulating the expression of the analogous enzyme-encoding genes studied, as well as collecting information on the subcellular locations of these enzymes. The results of comparative analysis of expression profiles, miRNA interaction networks, and subcellular localization seem to refute the hypothesis about functional redundancy since there are different regulators, subcellular locations and alternation in the expression pattern of the genes encoding analogous enzymes catalyzing a specific reaction.

ÍNDICE

ÍNDICE DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

1.1 Proteínas

Proteínas são macromoléculas biológicas, que participam de uma grande gama de reações ou processos nos organismos, tais como: transporte de substâncias dentro da célula, regulação da concentração de metabólitos, controle da expressão gênica, fluxo de substâncias através da membrana celular, formar o citoesqueleto, entre outras. Essas moléculas produtos da expressão gênica (1).

As proteínas podem ser caracterizadas por três propriedades: i) constituição de aminoácidos que compõem a sequência; ii) estrutura tridimensional (3D), que corresponde ao arranjo espacial da proteína; iii) função ou o fenótipo em uma determinada condição ou no organismo como um todo (2).

Os aminoácidos, unidades formadoras das proteínas, são moléculas orgânicas ligadas covalentemente entre si. Na natureza, 20 aminoácidos diferentes compõem as proteínas, sendo todos eles compostos de um carbono central (C_α), no qual estão ligados um átomo de hidrogênio (H), um grupo amino (NH_2), um grupo carboxil (COOH) e um grupo variável (grupo R). Os aminoácidos diferem entre si pela cadeia lateral (grupo R), também ligado ao Cα. A Figura 1.1 exemplifica uma das classificações feitas para os aminoácidos de acordo com as propriedades químicas do grupo R. A sequência de aminoácidos, também chamada de estrutura primária da proteína, é o arranjo linear dos resíduos de aminoácidos, unidos através de ligações peptídicas, que formam a cadeia polipeptídica (1,3). As sequências de aminoácidos das proteínas são definidas a partir da informação genética. Essa informação genética está contida na sequência de nucleotídeos do DNA que será transcrito em RNA mensageiro que ao ser traduzido determinará a sequência de aminoácidos da

proteína. Cada um dos 20 aminoácidos é codificado por um ou mais códons (trincas de nucleotídeos). A sequência de aminoácidos é a ligação entre a informação genética e a estrutura tridimensional da proteína, que desempenha a função biológica.

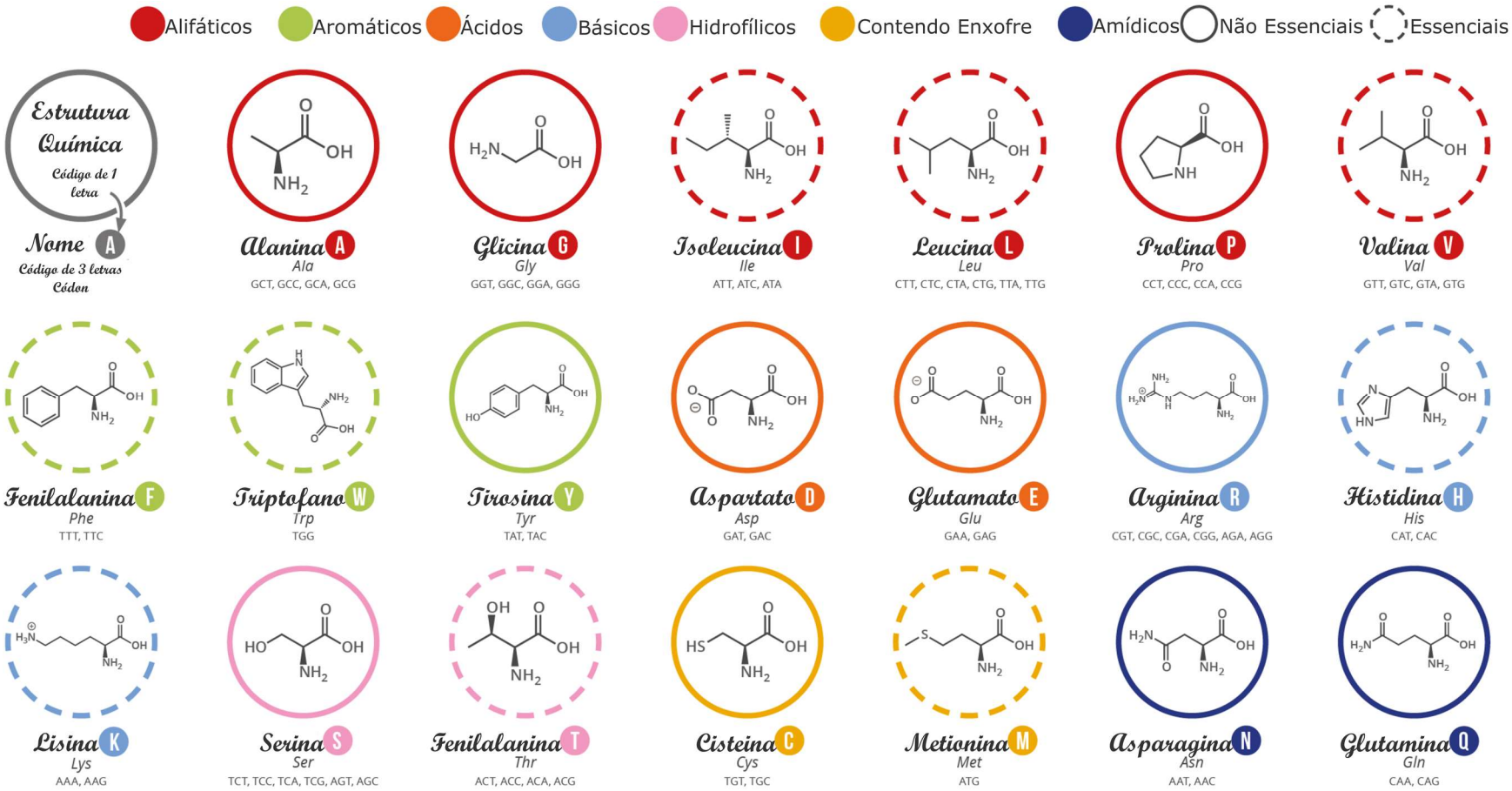


Figura 1.1 Representação dos 20 aminoácidos principais. As estruturas estão contidas nos círculos, com os nomes dos aminoácidos na parte inferior, seguidos da letra que representa o aminoácido, a abreviação e os códons que codificam o aminoácido. Em linhas pontilhadas estão os aminoácidos essenciais, aqueles que organismo não consegue sintetizar. Em linhas contínuas estão os aminoácidos não essenciais, que são aminoácidos que o organismo é capaz de produzir. Vermelho: aminoácidos alifáticos; Verde: aminoácidos aromáticos; Laranja: aminoácidos ácidos; Ciano: aminoácidos básicos; Rosa: aminoácidos hidrofílicos; Azul: aminoácidos amídicos; Amarelo: aminoácidos contendo enxofre. Modificado de: http://www.compoundchem.com/wp-content/uploads/2014/09/20-Common-Amino-Acids-v3.png

A partir da estrutura primária da proteína surgem os elementos de estrutura secundária (sendo as hélices α, as folhas β e voltas β as mais conhecidas), que corresponde ao segundo nível de organização de uma proteína e se refere ao arranjo espacial dos átomos na cadeia principal da cadeia polipeptídica, sem considerar a conformação de suas cadeias laterais ou sua relação com outros segmentos (1). O conhecimento da estrutura secundária pode auxiliar no reconhecimento de possíveis enovelamentos (“folds”), assim como na modelagem comparativa e por primeiros princípios (“ab initio”), nas quais a estrutura 3D de um polipeptídio é predita computacionalmente com auxílio de uma molécula similar, elucidada experimentalmente, que é usada como molde ou apenas a partir das características intrínsecas da própria sequência primária de aminoácidos, respectivamente (4). O termo enovelamento corresponde à combinação dos elementos de estrutura secundária em padrões conservados, moldados durante a evolução e que podem estar associados com uma determinada função (1,4,5) (Figura 1.2).

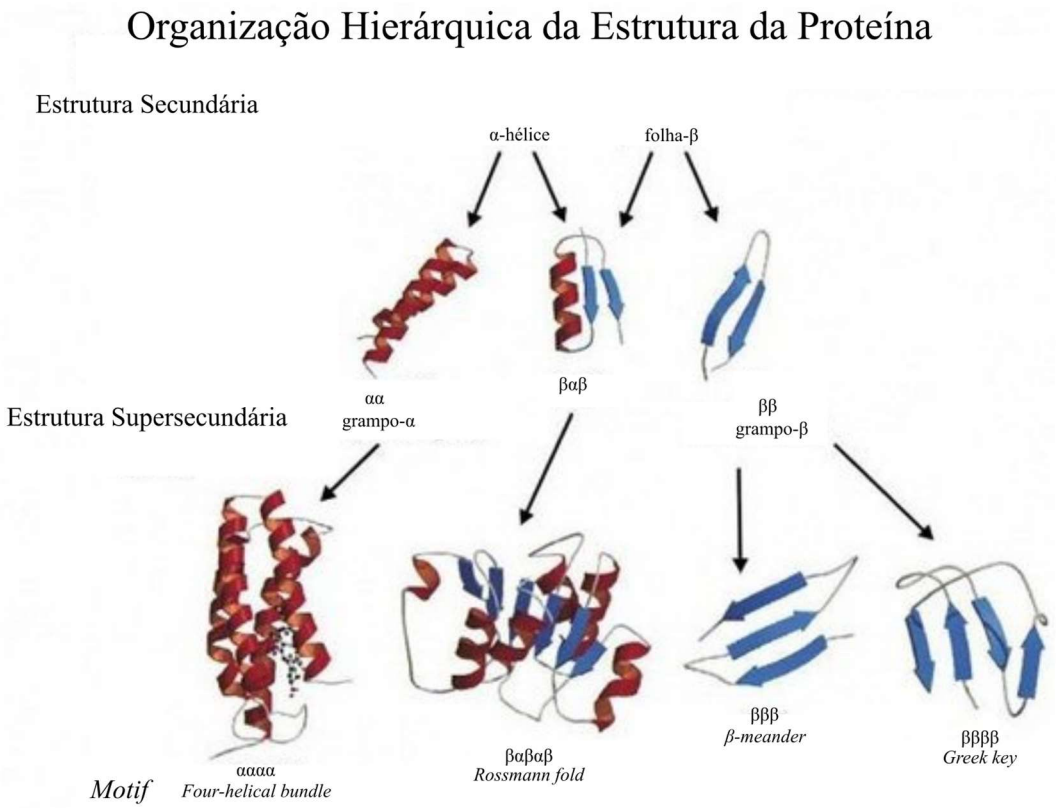


Figura 1.2 Organização hierárquica da estrutura proteica. Está representada a formação de quatro tipos de *folds* a partir dos elementos de estrutura secundária alfa-hélice e folha beta. Essas estruturas ao se combinarem geram padrões (superestrutura secundária) e, ao serem combinados entre si ou com outros padrões produzem as assinaturas estruturais das proteínas. Modificado de: https://qph.ec.quoracdn.net/main-qimg-899a8dbc47bedcdcf9700a8c98b28941-c).

O terceiro nível de organização corresponde ao arranjo espacial dos átomos que compõem a molécula. Essa organização 3D da proteína é sustentada por múltiplas ligações não covalentes e não covalentes entre as cadeias laterais dos aminoácidos contidos em sua sequência (1). Nesse nível de organização os elementos de estrutura secundária se dobram e se organizam até que a proteína atinja a sua conformação final, na qual a proteína assumirá a sua função biológica. Cada proteína apresenta uma estrutura terciária característica. Algumas proteínas apresentam um quarto nível de organização, denominada estrutura quaternária, na qual uma proteína contém duas ou mais cadeias polipeptídicas iguais ou distintas. O conhecimento sobre a arquitetura 3D das proteínas é vital para o perfeito entendimento de sua funcionalidade (4).

A necessidade de comparar estruturas 3D de proteínas resultou em métodos que estimam as distâncias entre os átomos das estruturas sobrepostas (6). A métrica mais comumente empregada em comparações estruturais é o RMSD (*Root-mean-square deviation*), que é calculada após a sobreposição das estruturas (6,7) (Equação 1.1). Nesse método, uma sobreposição perfeita entre as estruturas comparadas corresponde ao valor 0, enquanto valores maiores do que 0 refletem diferenças nas posições dos átomos no espaço tridimensional. No entanto, além do RMSD depender das distâncias atômicas existentes entre as estruturas comparadas, por empregar valores médios, todas as distâncias entre os resíduos são computadas igualmente, além de ser influenciado por modificações em trechos pontuais das estruturas, que tendem a elevar o valor de RMSD, mesmo que as duas estruturas comparadas sejam globalmente similares (7).

$$\begin{aligned}\mathrm{RMSD}(\mathbf{v},\mathbf{w}) &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|v_i - w_i\|^2} \\ &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}((v_{ix}-w_{ix})^2 + (v_{iy}-w_{iy})^2 + (v_{iz}-w_{iz})^2)}\end{aligned}$$

Equação 1.1 Equação de cálculo do RMSD para a determinação das distâncias atômicas entre pares de estruturas. n: número total de átomos presentes na comparação; x, y e z: coordenadas espaciais dos átomos presentes no arquivo PDB; v e w: representam os dois conjuntos de dados (as duas estruturas); i: representa um determinado átomo. A fórmula calcula o somatório dos desvios ao quadrado das coordenadas espaciais das moléculas comparadas. Esse valor é dividido pelo número de átomos e, em seguida, é obtida a raiz quadrada desse resultado.

Outra métrica que também usa a superposição de carbonos alfas de resíduos alinhados é o TM-score (*Template modeling score*), implementado no *software* TM-align (7). Essa métrica, derivada dos trabalhos de Levitt e Gerstein (8), é baseada na atribuição de pesos maiores aos pares de resíduos mais próximos em uma comparação estrutural. Diferente do RMSD, que não distingue variações locais e variações globais, o TM-score é mais sensível a variações estruturais ao longo de toda

a estrutura (globais), evidenciando a importância da conservação desses resíduos para as estruturas trabalhadas. O cálculo do TM-score está representado na Equação 1.2.

$$\text{TM-score} = \text{Max}\left[\frac{1}{L_N}\sum_{i=1}^{L_T}\frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}\right]$$

Equação 1.2 Equação utilizada para a determinação das similaridades entre pares de estruturas por TM-score. LN: tamanho da estrutura de referência; LT: tamanho do alinhamento entre os resíduos das duas cadeias; di: distância entre o ith par de resíduos alinhados; d0: escala para normalizar as diferenças de tamanho.

Por meio de programação dinâmica, o TM-align alinha os elementos de estrutura secundária dos pares comparados (resíduo a resíduo). Então é construída uma sobreposição (ótima) das representações 3D, retornando os valores de RMSD e TM-score normalizado (tendo o tamanho de uma das duas proteínas como referência para a normalização ou o tamanho médio das mesmas). O valor de TM-score obtido estará na faixa entre 0 e 1, assumindo-se que os valores maiores ou iguais a 0,5 possuem o mesmo enovelamento nas classificações do SCOP (9), enquanto valores inferiores a 0,2 representam similaridades aleatórias entre os pares de estruturas (7). O SCOP (*The Structural Classification of Proteins*) é uma base de dados que provê uma descrição detalhada das relações de estrutura de proteínas conhecidas, classificando hierarquicamente essas proteínas em: família, superfamília, enovelamento comum e classes (10,11).

1.2 Enzimas

Uma classe especial de proteínas, denominada enzimas, é capaz de catalisar a conversão de substratos em seus respectivos produtos. Enzimas são catalisadores biológicos capazes de acelerarem enormemente a velocidade das reações bioquímicas de maneira muito específica (3). Algumas das reações catalisadas pelas enzimas exigem a presença de compostos chamados cofatores para que a reação catalítica ocorra (3). O que são cofatores podem ser orgânicos ou inorgânicos. Os inorgânicos são os íons metálicos, que podem atuar livremente, como magnésio ou zinco, ou agrupados com outros átomos, como ferro-enxofre. Os cofatores orgânicos englobam as vitaminas e seus derivados, ácido ascóbico, por exemplo, e os cofatores não-vitamínicos, como o grupamento Heme presente nas moléculas de hemoglobina e o glutationa (12–14).

A nomenclatura das enzimas era feita de acordo com as reações que catalisavam, normalmente, adicionando-se o sufixo ‘-ase’ ao nome do substrato ou um termo que caracterizasse a reação (15). No entanto, esse tipo de nomenclatura, sem padrões ou regras, levava a redundância, em que uma mesma enzima poderia ter vários nomes e um mesmo nome poderia ser utilizado por mais de uma enzima (15).

Assim como qualquer informação científica, os termos precisaram ser padronizados para evitar comparações errôneas. Foi então que, na década de 1950, surgiram esforços na tentativa de criar um sistema de classificação enzimática baseado na funcionalidade dessas macromoléculas. O primeiro sistema de classificação dividia as enzimas em três classes: I) transferases, hidrolases e oxirredutases; II) liases e sintases; III) racemases (um tipo de isomerase) (15). Essa primeira tentativa de classificar as enzimas em níveis hierárquicos levou a IUBMB

(*International Union of Biochemistry and Molecular Biology*) <http://www.chem.qmul.ac.uk/iubmb/enzyme/> a formar um Comitê de Nomenclatura e classificação das enzimas. Com base na reação catalisada pela enzima, atualmente é atribuído um número da Comissão de Enzimas (*Enzyme Comission number – EC number*). De acordo com uma classificação hierárquica, onde cada enzima recebe um número de 4 dígitos (Figura 1.3). O primeiro dígito descreve a reação química geral catalisada pela enzima (a classe da enzima); os dois números subsequentes têm diferentes significados dependendo da classe da enzima (Por exemplo, enzimas com a anotação EC 1.1.1.- oxidam grupamentos CH-OH e possuem NAD+ ou NADP+ como aceptores de elétrons, enquanto as ligases EC 6.1.1.- promovem ligações Carbono-Oxigênio durante a formação de Aminoacil-RNAt e compostos relacionados; O quarto dígito descreve a especificidade da reação, definindo o substrato específico/produto ou cofatores utilizados (15)). O nível mais inclusivo desse sistema é o primeiro, a classe enzimática. Ao todo, existem seis classes de reações enzimáticas: Oxidorredutase (1.-.-.-; onde o substrato é oxidado e doa hidrogênio), Transferase (2.-.-.-; transferência de um grupamento de um composto para outro), Hidrolase (3.-.-.-; clivagem hidrolítica), Liase (4.-.-.-; clivagem de C-C, C-O e C-N, além de outras ligações por eliminação de grupamentos, formando ligações duplas ou anéis, ou formando essas estruturas por meio da adição de grupos de ligações duplas), Isomerase (5.-.-.-; catalisam rearranjos geométricos ou estruturais dentro de uma molécula) e Ligase (6.-.-.-; catalisam a união de duas moléculas através da hidrólise de uma ligação difosfato do ATP ou de outro trifosfato).

Vale destacar que, como em outras áreas da ciência, a nomenclatura de enzimas está sujeita a erros e novas informações podem modificar classificações antigas. Em alguns desses casos, a descrição de novas subclasses pode fazer com que enzimas previamente anotadas com um determinado EC sejam atribuídas a um ou mais ECs

recém-criados, fazendo com que os ECs antigos não estejam mais associados às enzimas em questão. Esses “ECs vazios” são preservados dessa forma, pois um pressuposto desse sistema de classificação é que o EC não pode ser reutilizado (15). Além disso, enzimas cujas reações suportam muitos substratos são descritas de maneira genérica. Por exemplo, o EC 1.4.3.3 (*D-amino acid oxidase*) tem como substrato D-aminoácidos, porém não é especificado quais são utilizados ou não (15). Outras reações são muito específicas, levantando questões sobre se tais características estão presentes em outros representantes do mesmo EC (15). Finalmente, compostos intermediários (não liberados), não são considerados na reação, uma vez que o EC descreve a reação geral catalisada (15).

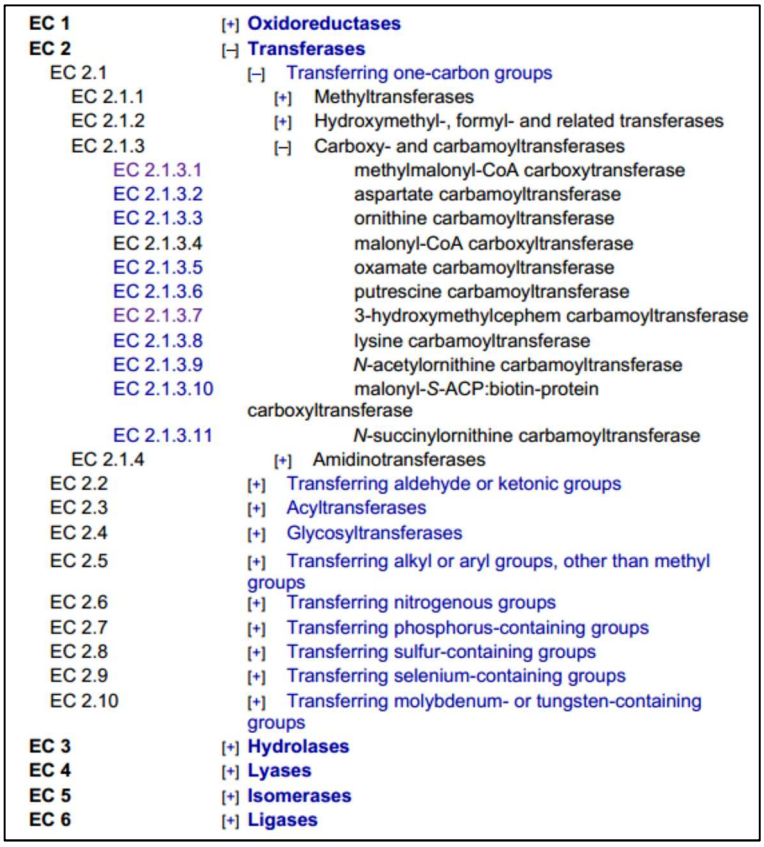


Figura 1.3 Sistema de classificação enzimática baseado na hierarquia funcional. Estão representadas as seis classes enzimáticas: oxirredutase, transferase, hidrolase, liase, isomarase e ligase. Estão exemplificadas algumas subclasses e reações enzimáticas presentes na classe transferase (2.-.-.-). (15).

As funções desempenhadas por uma enzima podem estar associadas à presença de domínios funcionais ao longo de sua sequência. Os domínios são trechos conservados evolutivamente entre grupos de sequências, podendo ocorrer isolados ou com outros domínios em uma mesma cadeia polipeptídica (16). Como os genes codificadores de proteínas podem passar por eventos de fissão/fusão gênica, a arquitetura de proteínas multidomínios pode variar ao longo da evolução, ocorrendo perda ou aquisição de domínios funcionais a partir de eventos de fissão ou fusão gênica, respectivamente, envolvendo genes codificadores de proteínas que podem ser evolutivamente relacionados ou não (17).

1.3 Convergência Evolutiva

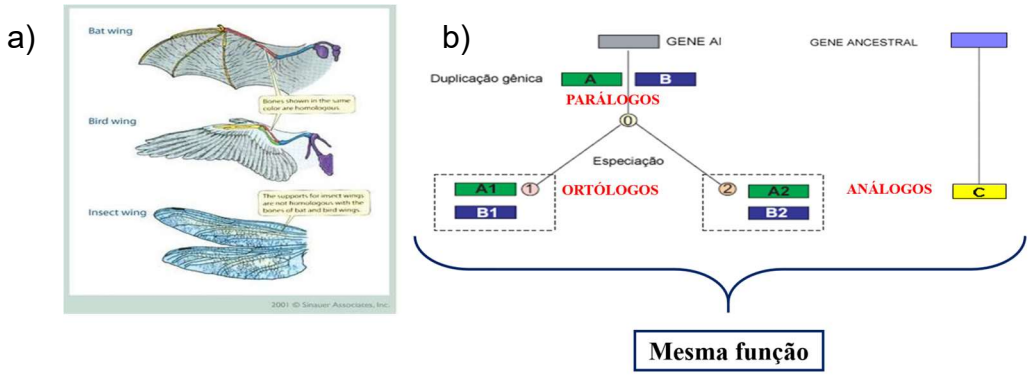


Figura 1.4. a) Analogia morfológica: estruturas morfológicas (asas) com origens embrionárias e evolutivas distintas, porém adaptadas ao voo; b) Analogia Molecular: os genes parálogos A e B, presentes nas espécies 1 e 2, descendem de uma sequência ancestral diferente da observada no gene C. Como os produtos dos genes A, B e C desempenham a mesma função, temos um caso de analogia funcional.

Convergência evolutiva é um processo que leva à similaridades que são produtos de causas não associadas a uma origem evolutiva comum (Figura 1.4) (18–20). Modificações que levam à tais similaridades podem ser reflexos de comportamentos decorrentes de adaptação em determinadas condições ambientais, como por exemplo, novos habitats ou novos padrões alimentares (21). A origem das asas, por exemplo, surgiu múltiplas vezes ao longo da árvore evolutiva da vida. Há 400 milhões de anos (Ma) essa característica surgiu e foi mantida em alguns representantes do grupo dos insetos, 200 Ma depois nas aves, há 230 Ma as asas foram desenvolvidas nos pterossauros – que, apesar de répteis voadores, não são ancestrais das aves – e 50 Ma nos morcegos (22,23).

Neste sentido, sabe-se que a compreensão das reações metabólicas é fundamental para entender a expressão fenotípica em todos os organismos vivos. Atualmente, o processo de reconstrução de redes metabólicas encontra-se bem estabelecido e tem sido aplicado a um número crescente de organismos, permitindo abordar uma variedade de questões científicas relevantes (24,25). Entretanto, ao

comparar vias bioquímicas preditas a partir da análise de genomas completamente sequenciados, inúmeras vias se apresentam incompletas ou mesmo ausentes em várias espécies (26–30). Em muitos casos, as enzimas desaparecidas foram substituídas por proteínas funcionalmente equivalentes, ou seja, capazes de catalisar as mesmas reações, mas exibindo virtualmente nenhuma similaridade ao nível de suas estruturas primárias e tampouco ao nível de suas estruturas terciárias. Algumas dessas enzimas "desaparecidas" foram identificadas e caracterizadas com algum detalhe (31,32). Aparentemente, enzimas análogas são frequentemente recrutadas a partir de superfamílias distintas (33,34), com algumas dessas formas alternativas compartilhando a reação catalisada e a configuração dos resíduos catalíticos (embora esses resíduos não compartilhem o mesmo enovelamento nestes casos) (32) (Figura 1.5).

Diferentes tipos de convergência ocorrem ao nível molecular e podem ser categorizadas em funcional, mecanística, estrutural e de sequência. Dessa forma, as enzimas cujas transformações químicas são definidas apenas por três dígitos da classificação por EC podem ter diferentes especificidades de reação (diferentes substratos/produtos ou cofatores), constituindo análogos mecanísticos, *i.e.*, enzimas não relacionadas que catalisam transformações químicas distintas através do mesmo mecanismo de ação (35,36). No entanto, este tipo de evento não é considerado neste trabalho, que se dedica exclusivamente a investigar analogia funcional no genoma humano.

É importante ressaltar ainda que investigações sobre a ocorrência de enzimas análogas em vias metabólicas podem não somente ampliar nossa compreensão sobre a origem e evolução das vias bioquímicas como também revelar novos alvos terapêuticos (37,38). Estudos *in silico* envolvendo identificação e anotação de enzimas análogas e reconstrução metabólica em diferentes organismos sugerem que estas

enzimas podem constituir um grande reservatório, ainda pouco explorado, de potenciais novos alvos para o desenvolvimento de fármacos dirigidos ao tratamento de doenças infecciosas (37,39,40). A abordagem empregada nestes trabalhos baseia-se no fato de que enzimas análogas funcionais presentes no hospedeiro e no patógeno possuem estruturas tridimensionais distintas, o que permite o desenvolvimento de inibidores específicos contra a enzima do patógeno, o que é um pré-requisito para o desenvolvimento racional de fármacos (38,41).

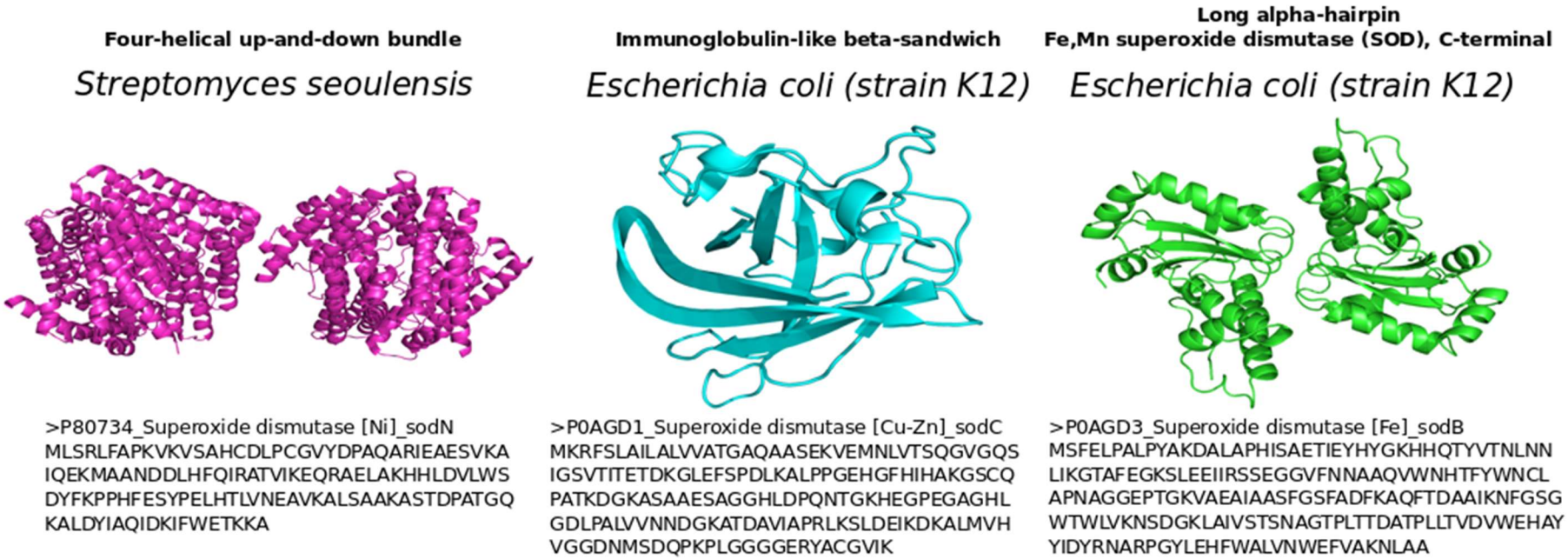


Figura 1.5 Representação de três enzimas que catalisam a a atividade superóxido dismutase (EC 1.15.1.1) codificadas nos genomas de duas bactérias distintas: *Streptomyces seoulensis* e *Escherichia coli* K-12. São constatadas diferenças nas estruturas primárias, estruturas 3D e nos enovelamentos (em negrito) entre todas enzimas. Neste trabalho, enzimas análogas pertencentes a espécies diferentes são denominadas intergenômicas, representadas nesta figura pelas enzimas sodN em *S. seoulensis* e sodC ou sodB em *E. coli*; por outro lado, as enzimas análogas sodC e sodB, codificadas no genoma de *E. coli*, são denominadas intragenômicas.

1.4 Genoma Humano

O conhecimento da sequência de ácidos nucleicos é de suma importância para as ciências biológicas (42). Isso resultou na iniciativa do Departamento de Energia Norte-Americano (DOE) de obter uma sequência genômica humana de referência, que, culminou, em 1990, no lançamento do conhecido (e ambicioso para a época) Projeto Genoma Humano (43). Onze anos mais tarde, como resultado desta iniciativa, os primeiros mapas completos do genoma humano foram publicados e disponibilizados para a comunidade científica (43,44). Outra consequência importante desta iniciativa foi que os avanços tecnológicos obtidos através do desenvolvimento deste projeto estimularam, durante os anos subsequentes, a criação e o desenvolvimento de inúmeros outros projetos genoma, levando à determinação do código genético de milhares de outros organismos, representando interesses médicos, comerciais, ambientais e industriais, ou contemplando organismos-modelos importantes para o desenvolvimento de pesquisas científicas (45). Além disso, o constante aprimoramento dos métodos de sequenciamento em larga escala, ocorrido nos últimos anos, tem aumentado substancialmente a rapidez e a eficiência com que genomas inteiros são sequenciados (46,47), estimulando cada vez mais diversas iniciativas no sentido de se obter sequências genômicas completas de um número crescente de organismos. As chamadas tecnologias de sequenciamento de alto rendimento têm permitido não apenas a caracterização de genomas inteiros, mas também a obtenção de perfis de transcrição completos, incluindo a quantificação de transcritos, entre outras possibilidades, de forma muito mais rápida e com menor custo, em relação às tecnologias anteriores.

Com a sequência completa do genoma humano em mãos, o grande desafio tornou-se interpretá-lo e aprender a usar essa informação para compreender melhor

a biologia humana. Embora exista uma gama de informações disponíveis sobre o papel funcional de muitas regiões identificadas tanto no genoma humano quanto em todos os outros genomas já sequenciados, especialmente em relação às regiões codificadoras de proteínas, inúmeras outras regiões permanecem sem função elucidada em todos esses organismos (48). Nesse contexto, com o objetivo de catalogar sistematicamente todos os elementos estruturais e funcionais codificados no genoma humano, o National Human Genome Research Institute (NHGRI) criou, em 2003, o consórcio internacional de pesquisa denominado ENCODE, Encyclopedia of DNA Elements (49). Desde sua criação, o consórcio tem feito um mapeamento exaustivo de regiões de transcrição, de associação com fatores de transcrição, de estrutura da cromatina e de modificação das histonas, permitindo a atribuição de função bioquímica para 80% do genoma humano, particularmente em regiões que não correspondem às tão bem estudadas regiões codificadoras de proteínas (50).

Transcriptoma corresponde ao conjunto completo de transcritos de RNA expressos em determinada condição, fase do desenvolvimento ou localização celular ou tecidual (51,52). A metodologia de sequenciamento do RNA (RNA-seq – *RNA sequencing*) vem sendo muito utilizada como uma forma de compreender o padrão de transcrição dos genes e sendo associada a estudos de associação de doenças com elementos funcionais dos genomas e o próprio funcionamento do genoma (52).

Por conta dos avanços decorrentes das tecnologias de sequenciamento de nova geração (NGS – *Next Generation sequencing*), RNA-seq vem substituindo outras tecnologias empregadas em análises transcriptômicas, como por exemplo, o microarranjo. O estudo de transcritos baseado na abordagem utilizando microarranjos possui a limitação, de avaliar apenas genes de sequências conhecidas, impossibilitando assim a descoberta de novos genes. A utilização de RNA-seq veio suprir essa limitação, uma vez que possibilita a detecção de transcritos

desconhecidos, o mapeamento em organismos sem genoma e transcritoma caracterizados, a detecção e diferenciação da expressão de alelos e isoformas, a detecção de limites de íntrons e éxons, entre outras. Uma grande vantagem da utilização de RNA-seq é a obtenção de dados relacionados aos transcritos oriundos do genoma total em uma determinada condição (53,54).

Exceto por Nanopore e PacBio, todas as abordagens seguem um fluxo metodológico que se inicia com a extração do RNA, seguida pela fragmentação desse RNA, a síntese do cDNA e a construção da biblioteca. Em seguida, milhões de sequências curtas chamadas leituras ou *reads* são obtidas; um controle de qualidade dos dados obtidos é realizado. Essas sequências curtas são alinhadas contra uma referência (genoma ou transcriptoma), mapeando os transcritos sequenciados em suas respectivas coordenadas, permitindo o conhecimento da estrutura e estimando os níveis de expressão de cada gene.

Frequentemente, essa quantidade de leituras pode variar entre amostras e bibliotecas de um mesmo experimento, afetando a interpretação dos dados. Dessa forma, a etapa de normalização permite que a expressão entre amostras seja comparável, reduzindo vieses causados pelo tamanho dos produtos gênicos e, consequentemente, a quantidade de leituras mapeadas por transcrito (55–58).

Quando não existe um genoma ou trecho de DNA de referência, os *reads* passam por uma montagem *de novo* para reconstruir transcritos de uma ampla faixa de níveis de expressão, semelhante à montagem de genoma. Após essa etapa, é feito o reconhecimento das regiões genômicas com leituras alinhadas (regiões codificadoras/éxons/junções), quantificado o número de leituras por região genômica e, posteriormente, a análise de expressão (53).

Neste contexto, análises particularmente interessantes abrangem o estudo de microRNAs, que são moléculas endógenas de RNA fita simples, com tamanho na

faixa de 18-22 nucleotídeos, envolvidas no silenciamento transcricional e pós-transcricional de genes (59–61). Todos os miRNAs são expressos em tipos específicos de células e tecidos e todos os microRNAs possuem uma região próxima da extremidade 5' conhecida como "*seed*", que interage com o RNAm via complementaridade de bases (62). Com exceção dos primeiros microRNAs, *let-7* e *lin-4*, esses reguladores recebem o prefixo 'miR' e um número sequencial característico, com miRNAs idênticos combinados no mesmo número, independente dos microRNAs agrupados pertencerem a genomas distintos ou não (63). São encontrados em plantas e animais (64) e estão envolvidos em vários processos biológicos. Alterações no perfil de expressão de microRNAs podem estar associados com algumas doenças (60,61). Mais de 1000 miRNAs são descritos no genoma humano e podem controlar a expressão de mais de 60% dos genes que compõem o genoma (61). Estes microRNAs estão contidos nos introns de genes codificadores de proteínas ou presentes como genes independentes (61).

1.5 Justificativa

Apesar de serem reconhecidas há muito tempo, embora erroneamente referidas na literatura científica mais antiga como isozimas (ou isoenzimas), isoformas, ou ainda enzimas classe/tipo I e classe/tipo II (por exemplo, (65), as enzimas funcionalmente análogas permanecem mal exploradas, e uma investigação abrangente sobre a ocorrência, distribuição e implicações de processos de convergência em atividades enzimáticas ainda não foi realizada. Questões fundamentais, tais como (i) de que forma enzimas análogas se originam, (ii) por que aparentemente tantos eventos de origem independente ocorreram durante a evolução, e (iii) quais as razões para a coexistência no mesmo organismo de formas enzimáticas distintas catalisando a

mesma reação bioquímica, permanecem sem resposta, bem como várias outras questões, por exemplo relacionadas à catálise de reações químicas semelhantes por enzimas com distintos arcabouços estruturais (31).

Surpreendentemente, inúmeras enzimas isofuncionais ainda não são reconhecidas como não-homólogas, apesar de evidências substanciais indicarem diferentes historias evolutivas (por exemplo, (34)). No entanto, em alguns desses casos não reconhecidos, demonstrou-se que as enzimas análogas analisadas possuíam uma insuspeitada história evolutiva distinta ou apresentavam características funcionais distintas (experimentalmente verificadas), como discutiremos mais adiante.

Assim, neste trabalho, propomos investigar o significado biológico da co-ocorrência de enzimas isofuncionais não-homólogas no metabolismo humano, caracterizando enzimas funcionais análogas identificadas em vias bioquímicas e processos anotados no genoma humano, bem como analisando seus perfis de expressão e regulação em diversos tecidos humanos normais.

Teoricamente, a expressão diferencial de duas ou mais formas análogas com a mesma atividade enzimática, em diferentes condições fisiológicas ou ambientais, poderia representar uma vantagem evolutiva, proporcionando certa flexibilidade metabólica e, portanto, uma maior capacidade adaptativa. Por possuírem estruturas tridimensionais distintas, é possível imaginar que estas formas alternativas exibam variações em diversas características funcionalmente importantes, como em sua afinidade pelo substrato, na utilização de cofatores, em sua meia-vida e estabilidade (tanto da proteína como de seu mRNA), em sua cinética reacional, etc. Estas diferenças poderiam ser exploradas pelos organismos que, dessa forma, poderiam obter um controle mais fino do seu metabolismo.

Portanto, nossa hipótese é que a coexistência de múltiplas formas enzimáticas não deve ser interpretada como redundância funcional. Em vez disso, essas formas enzimáticas podem estar envolvidas em papéis biológicos distintos (e provavelmente relevantes).

2 OBJETIVOS

2.1 Objetivo Geral

Investigar o papel biológico e evolutivo da existência de análogos funcionais intragenômicos em vias metabólicas na espécie humana.

2.2 Objetivos Específicos

(1) Identificar genes no genoma humano que codificam enzimas anotadas em atividades enzimáticas nas quais existem evidências de formas enzimáticas distintas (*i.e.*, com origem evolutiva e estruturas tridimensionais não relacionadas) catalisando a mesma reação bioquímica (analogia funcional intragenômica);

(2) Mapear as enzimas análogas intragenômicas preditas nas distintas atividades enzimáticas e vias bioquímicas caracterizadas no metabolismo humano;

(3) Reconhecer a estrutura e organização dos genes que codificam os pares ou grupos de formas enzimáticas análogas identificadas no genoma humano;

(4) Analisar comparativamente o perfil de transcrição dos genes que codificam os pares ou grupos de formas enzimáticas análogas identificadas.

3 MATERIAL E MÉTODOS

3.1 Predição Computacional de Formas Análogas

Sequências enzimáticas de 2.494 genomas compreendendo organismos dos três domínios da vida foram obtidos da base de dados KEGG versão 73.1 (66) <http://www.genome.jp/kegg/> e agrupados por atividade enzimática, com base no grau de similaridade entre suas sequências de aminoácidos, aplicando a metodologia desenvolvida por nosso grupo, implementada no *pipeline* AnEnPi (37), na qual sequências compartilhando a mesma anotação de atividade catalítica atribuídas a dois ou mais grupos distintos são consideradas possíveis casos de análogos funcionais, indicando que um ou mais eventos de origem independente ocorreram na atividade enzimática em questão ao longo da evolução. A análise envolvendo todos os genomas presentes no KEGG possibilita o reconhecimento de erros de anotação nesses dados, como uma forma enzimática formada por uma única sequência. Os agrupamentos utilizados nesta tese são oriundos do processamento feito por Alexander Franca Fernandes durante sua dissertação de mestrado (67).

Dessa forma, 1.159.633 sequências enzimáticas foram comparadas, separadamente por atividade enzimática (EC), todas contra todas, usando BLAST+ versão 2.2.30 (68) com os parâmetros *default*. Em seguida, o resultado do alinhamento foi transformado em um grafo, no qual cada sequência enzimática representa um nó. Para cada atividade enzimática, qualquer par de sequências (nó) que obtenha um *score* de alinhamento ≥ 120 é conectado por uma aresta; sequências ligadas dessa forma são presumidamente homólogas, sendo reunidas no mesmo grupo do AnEnPi; por outro lado, sequências de uma mesma atividade enzimática reunidas em grupos distintos do AnEnPi são presumidamente análogas. O número de

subgrafos obtidos representa o número de eventos de origem independente ocorrido em cada atividade enzimática, ou seja, o número de vezes que uma determinada atividade enzimática surgiu *de novo* ao longo da evolução. O valor de corte de similaridade usado na fase de agrupamento (BLAST *score* ≥ 120) é baseado em uma significante observação experimental: enzimas que comprovadamente compartilham a mesma atividade enzimática e apresentam diferenças significativas em suas estruturais 3D (com base em alinhamentos estruturais), possuem *score* inferior a 120 quando suas sequências de aminoácidos são comparadas com BLAST (33). Apesar da ausência de similaridade de sequência detectável ser frequentemente atribuída à divergência entre sequências homólogas durante a evolução, é observado que muitas formas alternativas de enzimas que catalisam a mesma reação bioquímica apresentam estruturas 3D significativamente distintas e, por esse motivo, possivelmente evoluíram independentemente (33,34).

Em seguida, o resultado do AnEnPi foi processado da seguinte forma: (i) ECs incompletos, com todas as suas sequências enzimáticas, foram removidos. Enzimas cujas transformações químicas estivessem definidas apenas até o terceiro dígito da classificação EC podem ter diferentes especificidades de reação (diferentes substratos/produtos ou cofatores) e, neste caso, os análogos preditos poderiam corresponder a um tipo de analogia conhecida como mecanística (35,36). Entretanto, esse tipo de analogia não é parte do escopo deste trabalho, o qual se dedica exclusivamente em estudar analogia funcional; (ii) atividades enzimáticas nas quais as sequências foram anotadas como "subunidade" e "cadeia" foram manualmente inspecionadas e excluídas, pois a presença de enzimas heteromultiméricas no conjunto de dados pode inflar o número de eventos de analogia detectados. Esse problema surge durante o processo de anotação das sequências enzimáticas, no qual diferentes subunidades (ou cadeias) de uma enzima multimérica frequentemente

herdam a atividade anotada para a enzima como um todo, desconsiderando suas origens evolutivas e a participação na atividade em questão; (iii) atividades enzimáticas contendo grupos formados por uma única sequência humana foram removidas. No caso de apenas uma única sequência ser diferente de dezenas ou centenas de outras sequências enzimáticas (de humanos e/ou outras espécies), isso pode estar relacionado a um erro de anotação funcional; (iv) atividades enzimáticas sem a ocorrência de formas alternativas (compostas por um único grupo do AnEnPi) foram removidas.

3.2 Validação da Analogia Intragenômica Predita em *Homo sapiens*

Foram usadas anotações de domínios, superfamília/enovelamento e estrutura tridimensional (3D) recuperadas para enzimas de um mesmo grupo e de grupos distintos de uma mesma atividade enzimática, com o objetivo de confirmar supostos casos de analogia detectada dentro do genoma humano (analogia intragenômica). Esses dados foram obtidos do Pfam 27.0 (69) <http://pfam.sanger.ac.uk/> e SUPERFAMILY 1.75 (70) <http://supfam.org/SUPERFAMILY>. Estruturas 3D de proteínas experimentalmente resolvidas foram obtidas da base de dados PDB <http://www.rcsb.org/> e informações prévias sobre convergência em atividades enzimáticas foram selecionadas a partir da literatura científica (34). Com base na classificação de superfamília, os resultados foram divididos em dois conjuntos de dados: (-) e (+). Atividades enzimáticas potencialmente catalisadas por formas proteicas pertencentes a pelo menos duas superfamílias distintas foram atribuídas ao conjunto de dados (+), caso contrário foram atribuídas ao conjunto de dados (-). Considerando o conjunto de dados (+), realizamos uma comparação entre todas as sequências classificadas como capazes de catalisarem uma mesma reação, usando

o algoritmo rigoroso de alinhamento global de sequências implementado no software Needle (71), produzindo o conjunto de dados *bona fide*.

Para sequências sem informação 3D foram gerados modelos estruturais por meio do software de modelagem comparativa Modeller (72). Para isso, foram recuperados moldes estruturais a partir da base de dados PDB, através de buscas por similaridade com o BLAST (68) (cobertura em relação à sequência de entrada > 70%; cobertura em relação à sequência de pesquisada > 90%; identidade > 30%; e-value < 10-3). Foram gerados 50 modelos estruturais e o melhor modelo para cada proteína foi selecionado com base no menor DOPE *score* (*Discrete Optimized Protein Energy*). O DOPE score é um método estatístico empregado para avaliação da qualidade dos modelos, que é responsável pela forma finita e esférica das estruturas nativas.

Subsequentemente, os modelos com qualidade satisfatória foram avaliados com SAVES <http://services.mbi.ucla.edu/SAVES/> e MolProbity (73). As cadeias laterais nesses modelos foram ajustadas com KiNG (74), e a minimização de energia foi feita com ModRefiner (75). As representações 3D dessas estruturas foram geradas com PyMOL (The PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC) e o alinhamento estrutural das proteínas foi conduzido com TM-align (7). Os valores de RMSD e TM-scores (76) foram calculados com o pacote TM-align. Os valores de TM-score foram normalizados pelo tamanho médio das cadeias de cada estrutura envolvida nas comparações par a par.

3.3 Mapeamento Genômico e Metabólico de Enzimas Análogas

Com o intuito de conhecer o padrão de distribuição dos genes codificadores de enzimas relacionadas com analogia intragenômica em humanos, foi estudado o contexto genômico desses genes. Coordenadas genômicas dos genes codificadores

de formas alternativas no conjunto de dados *bona fide* (+) foram obtidas do Ensembl (versão do genoma: GRCh38) (77). O ideograma representando a localização cromossômica desses genes foi criado com o software PhenoGram <http://ritchielab.psu.edu/>. Além disso, um diagrama circular evidenciando as distâncias entre os genes codificadores de formas alternativas (distintos agrupamentos produzidos pelo AnEnPi para um mesmo EC), assim como genes codificadores de formas enzimáticas homólogas (ou seja, isoenzimas, enzimas humanas reunidas no mesmo grupo do AnEnPi) foi construído com o programa Circos (78).

3.4 Análise comparativa do perfil de transcrição dos genes que codificam as formas enzimáticas análogas identificadas

Para identificar e comparar os perfis de expressão dos genes codificadores do repertório de enzimas análogas identificadas no metabolismo humano, utilizamos dados de *counts* obtidos para 8.555 amostras de 53 tecidos humanos saudáveis disponíveis publicamente no portal GTEx (79,80) <https://www.gtexportal.org/home/>. Os dados de RNA-seq disponibilizados pela plataforma GTEx foram produzidos pela metodologia Illumina TruSeq, produzindo leituras *paired-end* com 76bp. Esse método resulta na seleção de *RNA-Seq polyA+* não-fita específicos.

Em seguida, esses dados os dados da foram alinhados contra o genoma humano de referência (HG19), por meio do software Tophat v1.4.1 (81), tendo o transcriptoma humano do GENCODE v19 também como referência, e combinando as isoformas existentes em uma única forma canônica. Os dados disponibilizados publicamente pelo consórcio foram selecionados com base na qualidade do alinhamento das leituras contra o genoma humano, utilizando o programa RNA-SeQC. Foram mantidas apenas

as leituras que: i) apresentavam o par correspondente no alinhamento (*paired-end*); ii) cujo o número de bases não alinhadas entre suas sequências e o genoma humano não era superior a 6; iii) estavam 100% contidas dentro dos limites dos éxons (leituras sobrepondo introns não foram contadas); iv) não apresentaram mapeamentos múltiplos contra o genoma humano, o que corresponde a qualidade de 255 no programa TopHat. Ao fim dessa etapa de preprocessamento, o programa RNA-SeQC foi utilizado pelo consórcio GTEx para contabilizar o número de leituras alinhadas por elemento genômico (*counts*).

A partir do arquivo contendo as informações sobre *counts* para os 53 tecidos, disponibilizado pelo consórcio GTEx, utilizamos esses valores de *counts* como entrada para as análises de expressão gênica dos genes relacionados com analogia intragenômica humana. Ao todo, as 8.555 amostras, de 53 tecidos diferentes foram obtidas da versão V6p dessa base de dados (Figura 3.1). O arquivo de anotação disponibilizado na página do consórcio foi processado utilizando *scripts* caseiros escritos em linguagem bash, com o objetivo de criar um arquivo *design* para as análises de expressão diferencial. O arquivo *design* relaciona o código da amostra com seu tecido biológico correspondente, por exemplo, o código GTEX-111CU-1826-SM-5GZYN que representa uma das 350 amostras de tecido adiposo subcutâneo. Dessa forma, foi possível trabalhar os dados no software R, com o pacote edgeR (82), disponibilizado na plataforma Bioconductor (83).

Além desse arquivo, também foi obtido um arquivo GTF (*General Transfer Format*) contendo as anotações de todos os dados trabalhados na plataforma (“gencode.v19.genes.V6p_model.patched_contigs”). Vinte e nove classes de elementos genômicos (*features*) foram observados nesse conjunto de dados. Esse arquivo foi modificado de forma a restar apenas informações sobre as coordenadas

inicial e final nos segmentos cromossômicos que contém cada um dos genes, assim como a fita na qual estão codificados (+ ou -) e o seus tamanhos (Tabela 3.1).

Tabela 3.1 Propriedades genômicas descritas no arquivo de anotação genômica disponibilizado na plataforma GTEx ("gencode.v19.genes.V6p_model.patched_contigs"). Para cada gene é indicado seu cromossomo, coordenada inicial, coordenada final, fita (positiva ou negativa) e tamanho do gene.

GENE	CROMOSSOMO	COORDENADA INICIAL	COORDENADA INICIAL	FITA	TAMANHO
DDX11L1	1	11869	14362	+	2493
WASH7P	1	14413	29553	-	15140
FAM138A	1	34554	36081	-	1527
OR4G4P	1	52473	54936	+	2463
OR4G11P	1	62948	63887	+	939
OR4F5	1	69091	70008	+	917

Após a criação/modificação desses dois arquivos, a análise de expressão gênica teve início. As tabelas com os valores de *counts*, *design* e anotação foram carregadas na plataforma R. A atribuição das amostras aos seus respectivos tecidos foi estabelecida através da associação dos identificadores das amostras presentes no arquivo com os tecidos com os mesmos identificadores no arquivo *design*.

Em seguida, o pacote edgeR foi carregado e uma lista foi construída a partir das informações da tabela de *counts*, da tabela com a anotação dos elementos genômicos (construída a partir do GTF) e dos grupos (tecidos/linhagens celulares). Com o objetivo de eliminar possíveis artefatos, genes com baixa representatividade foram descartados do conjunto de dados, mantendo-se apenas aqueles que apresentaram valor de CPM (*counts per million*) superior a 1 em pelo menos 50% das amostras trabalhadas (4.277 amostras).

Posteriormente, foi recalculado o tamanho da biblioteca e o fator de normalização (*calcNormFactors*). Esse fator de normalização minimiza as diferenças entre amostras para os genes estudados (*log-fold changes*). Nesses casos, um valor de fator de normalização abaixo de 1 indica que um subconjunto de genes foi sobre-amostrado durante o sequenciamento, fazendo com que os demais genes tenham uma quantidade de *counts* menor do que seria esperado, dado o tamanho da

biblioteca. Então, o tamanho da biblioteca é escalonado para baixo. Contrariamente, fatores com valores maiores do que 1 escalonam para cima o tamanho da biblioteca. Foi estimada também a dispersão dos elementos genômicos no conjunto de dados. O pacote edgeR usa a média de expressão como representação da abundância de cada gene em uma determinada amostra, permitindo calcular tanto o valor de dispersão que representa todo o conjunto de dados (*common.dispersion*), quanto um valor que caracterize os genes entre amostras (*tagwise.dispersion*).

Ao fim do *pipeline*, os valores de CPM normalizados foram exportados para uma tabela. A partir dessa tabela, foram selecionados os genes codificadores de enzimas isofuncionais não-homólogas identificadas no genoma humano e determinado o valor médio por tecido para cada gene.

Com base nessa lista de genes, foi construído um arquivo contendo os valores médios de CPM para cada uma das atividades enzimáticas com evidência de analogia intragenômica em humanos. Foram descartados das análises os ECs nos quais todos os genes eram representantes de uma única forma enzimática. Dessa forma, o perfil de expressão dos genes codificadores de formas análogas foi analisado através do programa Expander (84,85), aplicando o algoritmo de agrupamento hierárquico aglomerativo, que utiliza uma medida de distância, baseada no coeficiente de correlação de Pearson. O programa produz uma matriz com os níveis de expressão de cada gene nos diferentes tecidos estudados. Dessa forma, o padrão de expressão dos genes é comparado ao dos demais genes e as similaridades e diferenças são representadas na forma de um dendrograma, agrupando genes com perfis mais similares ao longo da árvore, além dos comprimentos dos ramos evidenciarem as distâncias entre os agrupamentos. Consequentemente, considerando cada gene, foi calculado o valor médio das distâncias entre os agrupamentos hierárquicos

compostos por formas análogas e entre parálogas (pertencentes à mesma forma enzimática) para cada atividade enzimática.

Finalmente, a partir das predições computacionais disponíveis na base de dados TargetScan v7 (86–89), identificamos microRNAs (miRNA) humanos que possivelmente modulam a expressão dos genes estudados. Além disso, também buscamos informações sobre as localizações subcelulares das enzimas codificadas por esses genes usando a plataforma The Human Protein Atlas <http://proteinatlas.org/> (90). Os mapas representando as redes de interação genes/miRNA foram gerados com uso do programa Cytoscape versão 3.6.0 (91).

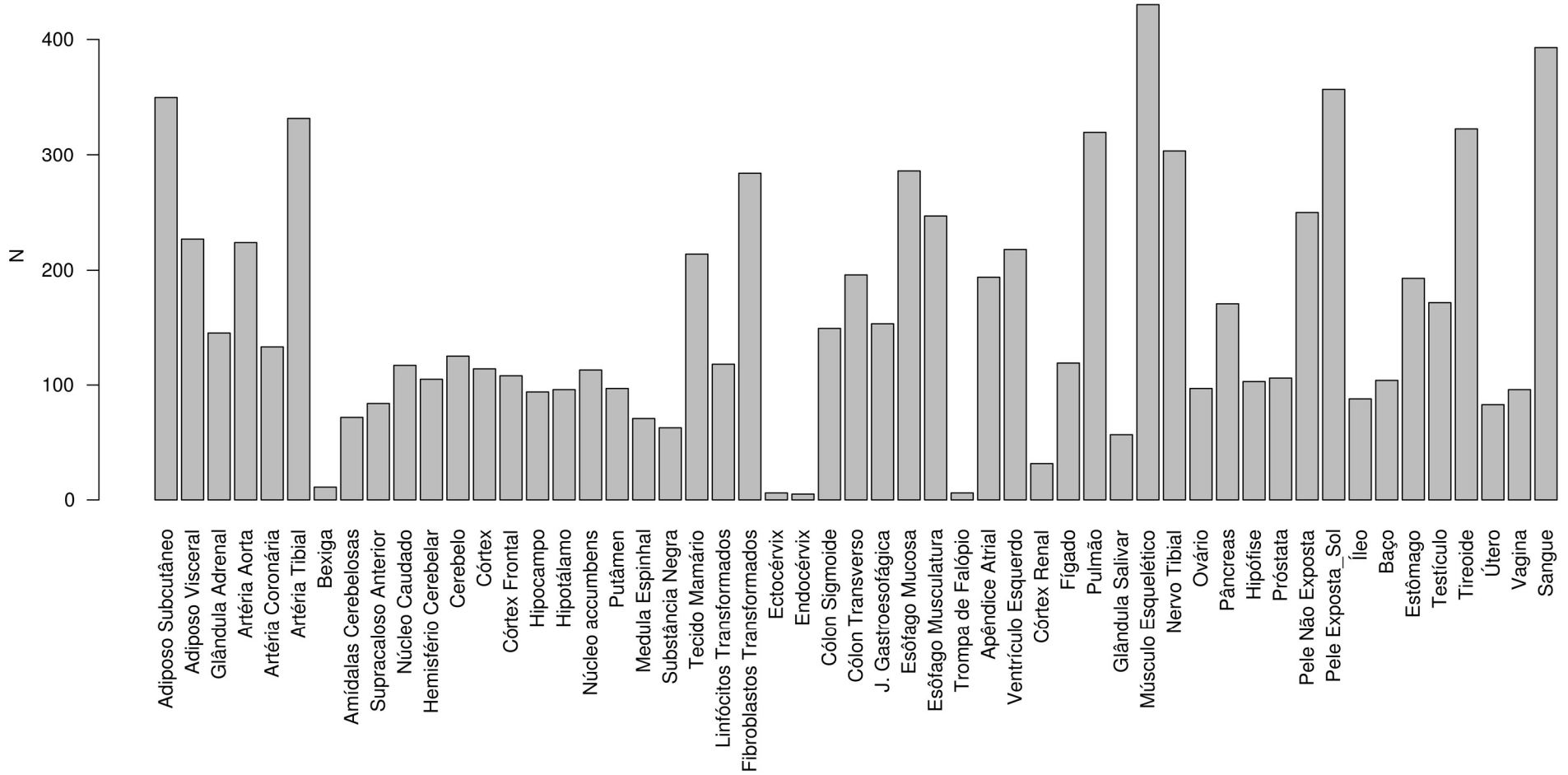


Figura 3.1 Amostras de RNA-Seq selecionadas no repositório GTEx. Estão representados os 53 tecidos e a distribuição das 8.555 amostras que compõem esse conjunto de dados. Eixo Y: frequência absoluta (número total) de amostras em cada tecido estudado; eixo X: representação dos 53 tecidos humanos que compõem o conjunto de dados.

4 RESULTADOS E DISCUSSÃO

4.1 Repertório de Enzimas Isofuncionais Não-Homólogas em Humanos

Uma fonte de informação sobre as atividades enzimáticas e as vias metabólicas é a base de dados KEGG Pathway disponível na plataforma KEGG (*Encyclopedia of Genes and Genomes*), que compreende uma coleção de mapas elaborados manualmente representando o conhecimento atual sobre redes de interação molecular em processos biológicos ou vias bioquímicas. Assim, a partir da base de dados KEGG versão 73.1, obtivemos 1.159.633 sequências enzimáticas codificadas em 2.494 genomas, abrangendo os três domínios de vida (*Archaea*, *Bacteria* e *Eukarya*), distribuídas em 3.825 atividades enzimáticas. Entre essas atividades enzimáticas, 3.572 estavam anotadas com os quatro dígitos da classificação por EC, contendo 1.025.885 sequências enzimáticas. Por outro lado, 253 ECs incompletos foram identificados (definidos até o primeiro, segundo ou terceiro dígito do sistema de classificação por EC), contendo 133.748 sequências.

A predição computacional com AnEnPi (incluindo todos os organismos e atividades enzimáticas disponíveis na base de dados KEGG) resultou em 2.203 atividades enzimáticas nas quais as sequências enzimáticas foram separadas em dois ou mais grupos distintos, compreendendo 1.996 atividades enzimáticas com anotação de quatro dígitos do EC. Considerando apenas a inferência de convergência em atividades enzimáticas anotadas com quatro dígitos pela classificação EC no genoma humano, encontramos 150 ECs (2.288 sequências de proteínas) identificadas pelo AnEnPi que sustentam eventos de origem evolutiva *de novo*. Após a remoção do nosso conjunto de dados de atividades enzimáticas nas quais uma ou mais sequências enzimáticas foram anotadas como "subunidades" e "cadeias", bem como

de atividades enzimáticas compostas de grupos formados por uma única sequência humana e nas quais todas as sequências enzimáticas foram reunidas em um único grupo (enzimas isofuncionais homólogas, portanto), obtivemos 116 sequências enzimáticas pertencentes a 32 atividades enzimáticas distintas do metabolismo humano.

Um fluxograma representando as análises subsequentes realizadas com os dados é mostrado na Figura 4.1. De forma geral, 116 sequências enzimáticas foram inicialmente preditas como pares ou grupos de formas alternativas em 32 atividades enzimáticas do metabolismo humano. Destas, 70 sequências enzimáticas, compreendendo 15 ECs, foram atribuídas ao conjunto de dados (+), no qual todas as atividades enzimáticas são compostas de formas enzimáticas alternativas putativas pertencentes a pelo menos duas superfamílias distintas. As 46 sequências restantes (pertencentes a 17 ECs), atribuídas ao conjunto de dados (-), foram rejeitadas das nossas análises.

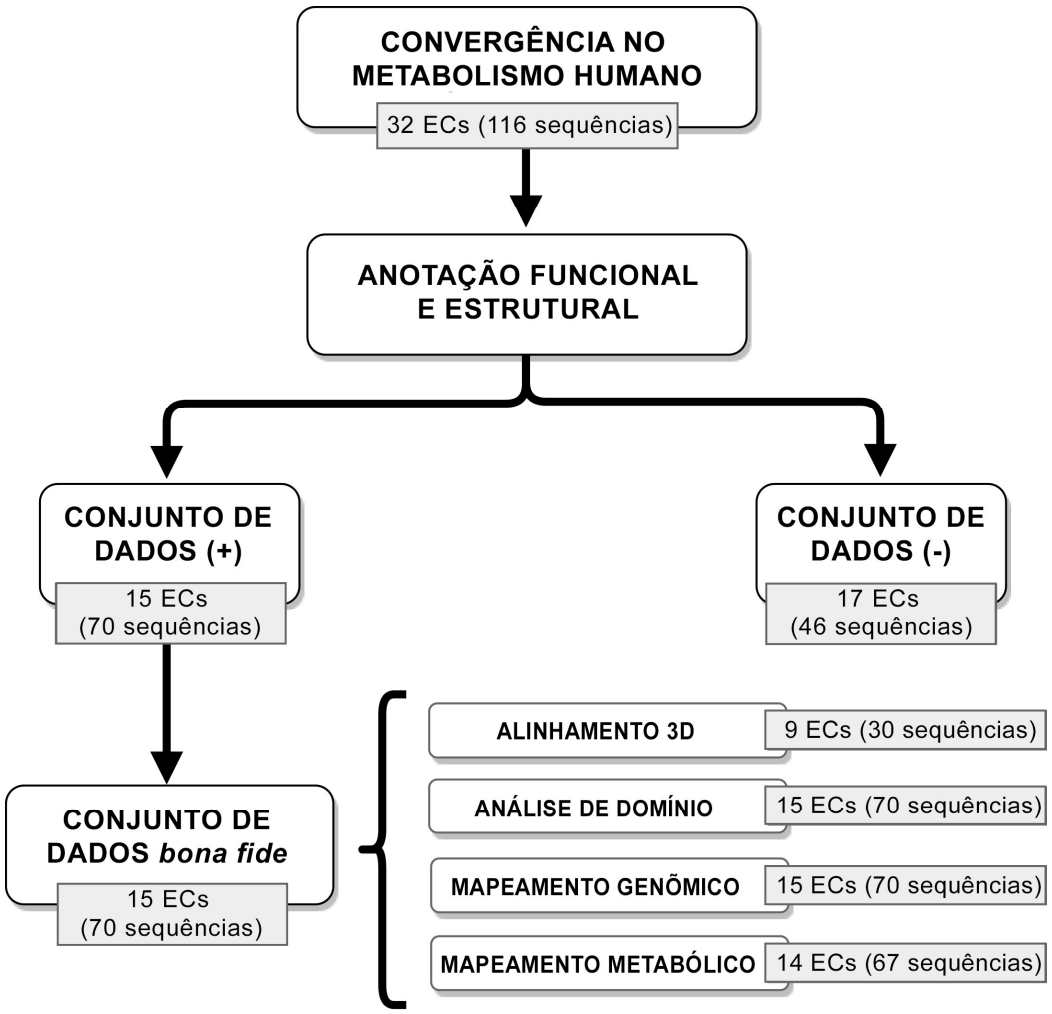


Figura 4.1. Esquema representando nossa metodologia para a identificação de análogos funcionais intragenômica no metabolismo humano.

A classe Hidrolase (36 sequências em 6 ECs) foi a classe mais frequente no conjunto de dados (+), seguida por Transferase (17 sequências em 4 ECs), Oxidoredutase (6 sequências em 2 ECs), Liase (7 sequências em 1 EC) e Isomerase (4 sequências em 2 ECs). Não foram encontradas evidências de convergência na classe Ligase, pois as todas formas enzimáticas recuperadas pelo AnEnPi apresentaram a anotação como “cadeia” ou subunidade”. Por outro lado, essas 15 atividades enzimáticas encontram-se mapeadas em 45 vias ou processos bioquímicos de várias das maiores classes metabólicas: Envelhecimento, Câncer, Metabolismo de carboidratos, Comunidade celular em eucariotos, Desenvolvimento, Sistema digestivo, Sistema endócrino, Biossíntese e metabolismo de glicanos, Doenças

imunes, Metabolismo de lipídios, Metabolismo de cofatores e vitaminas, Metabolismo de outros aminoácidos, Doenças neurodegenerativas, Metabolismo de nucleotídeos, Replicação e reparo, Transdução de sinal, Catabolismo e transporte, Biodegradação e Metabolismo de xenobióticos (Anexo I).

Vale ressaltar que 12 dessas 15 atividades enzimáticas do conjunto de dados (+) (~73%) foram previamente descritas como possuindo evidências de analogia (28): 1.3.1.20 (*Trans-1,2-dihydrobenzene-1,2-diol dehydrogenase*), 1.15.1.1 (*Superoxide dismutase*), 2.7.4.21 (*Inositol-hexakisphosphate kinase*), 3.1.1.3 (*Triacylglycerol lipase*), 3.1.1.29 (*Aminoacyl-tRNA hydrolase*), 3.1.2.2 (*Palmitoyl-CoA hydrolase*), 3.1.3.2 (*Acid phosphatase*), 3.1.3.5 (*5'-nucleotidase*), 3.1.4.12 (*Sphingomyelin phosphodiesterase*), 4.2.99.18 (*DNA-(apurinic or apyrimidinic site) lyase*), 5.3.99.2 (*Prostaglandin-D synthase*) e 5.3.99.3 (*Prostaglandin-E synthase*).

O banco de dados SUPERFAMILY (70) consiste de uma coleção de perfis HMM (*hidden Markov models*), representando domínios estruturais proteicos, segundo a classificação do SCOP. Consequentemente, uma superfamília agrupa domínios com parentesco evolutivo. Portanto, considerando a classificação do SUPERFAMILY, foram identificadas 39 superfamílias distintas (38 enovelamentos diferentes) entre os casos putativos de enzimas análogas no conjunto de dados (+) *bona fide*. As mais frequentes superfamílias são: *alpha/beta-Hydrolases* (13), *Phosphoglycerate mutase-like* (6), *Lipase/lipooxygenase domain* (*PLAT/LH2 domain*) (6), *ADP-ribosylation* (5), *DNase I-like* (4), *HAD-like* (4), *Metallo-dependent phosphatases* (3), *SAICAR synthase-like* (3), *NAD(P)-linked oxidoreductase* (2), *Protein kinase-like* (*PK-like*) (2), *Cu,Zn superoxide dismutase-like* (2), *DNA-glycosylase* (2), *Glutathione synthetase ATP-binding domain-like* (2), *GST C-terminal domain-like* (2), *Nucleotide-diphospho-sugar transferases* (2), *S13-like H2TH domain* (2), e *Thioredoxin-like* (2), seguidas por 22 superfamílias diferentes representadas uma única vez. Por outro lado, foram

observados 51 domínios/famílias distintas do Pfam nessas 70 enzimas. Dessas 70 sequências, 29 enzimas possuem arquitetura multidomínio e 41 são compostas por (ou anotadas como) domínio único. Três domínios são compartilhados entre determinadas atividades enzimáticas: *His_Phos_2* (ECs 2.7.4.21 e 3.1.3.2), *Metallophos* (ECs 3.1.3.2, 3.1.3.5 e 3.1.4.12), e *Exo_endo_phos* (ECs 3.1.4.12 e 4.2.99.18). Com duas exceções, formas enzimáticas atribuídas a grupos distintos possuem diferentes composições de domínios, indicando que, dentro do mesmo agrupamento (para um determinado EC), as sequências possivelmente compartilhem uma origem comum. Entretanto, formas alternativas da atividade enzimática 2.7.1.67 apresentam a mesma composição de domínio (*PI3_PI4_kinase*), embora sejam membros de superfamílias não relacionadas (*ARM repeat*, *Protein kinase-like (PK-like)* e *ADP-ribosylation*). Já a atividade enzimática 4.2.99.18, por outro lado, exibe um padrão muito mais complexo de composição de domínios e superfamília (Anexo I).

Para medir a similaridade entre essas 70 sequências no conjunto de dados (+), realizamos um alinhamento rigoroso global par a par de sequências (Tabela 4.1). Os maiores valores de *score*, similaridade e de identidade foram observados entre as formas enzimáticas pertencentes ao mesmo grupo gerado pelo AnEnPi, como era esperado, uma vez que as enzimas que compartilham a mesma atividade enzimática, reunidas no mesmo grupo, são presumidamente homólogas. Resultados semelhantes foram obtidos quando as estruturas 3D destas sequências enzimáticas foram comparadas utilizando o TM-score (68) e RMSD para estimar a similaridade entre elas. Aplicamos os seguintes limiares para distinguir estruturas relacionadas e não relacionadas: TM-score < 0.2, indicando uma provável origem evolutiva distinta, e TM-score > 0.5, possivelmente correspondendo ao mesmo enovelamento no SCOP (9) ou CATH (92). A maioria das formas alternativas obteve TM-scores < 0.5 quando suas estruturas foram alinhadas (Tabela 4.1 e Anexo II). Portanto, as comparações entre

as sequências pertencentes ao mesmo grupo AnEnPi/EC resultaram em valores de RMSD tendendo a zero e TM-scores perto de 1, indicando uma possível origem evolutiva comum. Quando se compararam sequências pertencentes a grupos distintos do mesmo EC, observou-se uma tendência inversa nesses resultados, também conforme esperado (Tabela 4.1).

Os valores intermediários de *score* observados entre o produto do gene NT5C3A e as formas alternativas codificadas pelos genes NT5C (0.50279) e NT5M (0.50929), bem como entre os produtos dos genes CEL e LIPF (0.47684), podem ser atribuídas aos enovelamentos *HAD-like* e *alfa/beta-hidrolases* compartilhados entre eles, respectivamente.

Tabela 4.1. Similaridade estrutural e perfis de similaridades entre sequências contidas em cada atividade enzimática pertencente ao conjunto de dados (+) de enzimas análogas intragenômicas. O número do cluster apresentado na tabela é o mesmo produzido pelo AnEnPi, ou seja, com base na análise envolvendo todos os genomas. Os clusters representados na tabela e os demais dados da tese correspondem aos agrupamentos com sequências de *H. sapiens*.

EC	Gene[1]	Uniprot[1]	PDB[1]	Cluster[1]	Gene[2]	Uniprot[2]	PDB[2]	Cluster[2]	Identidade	Similaridade	*Score*	TM-Score	RMSD
1.3.1.20	DHDH	Q9UQ10	2O48*	1	AKR1C2	P52895	2HDJ	2	16.3%	28.0%	40.0	0.34186	6.09
	DHDH	Q9UQ10	2O48*	1	AKR1C1	Q04828	1J96	2	16.0%	27.5%	33.0	0.33888	5.89
	AKR1C2	P52895	2HDJ	2	AKR1C1	Q04828	1J96	2	97.8%	98.5%	1662.0	0.99487	0.40
1.15.1.1	SOD2	P04179	1LUV	1	SOD1	P00441	4XCR	2	13.6%	24.8%	37.5	0.24203	4.48
	SOD2	P04179	1LUV	1	SOD3	P08294	2JLP	2	3.7%	6.7%	21.5	0.28352	5.09
	SOD1	P00441	4XCR	2	SOD3	P08294	2JLP	2	25.1%	34.4%	265.5	0.70353	1.76
2.4.1.22	B4GALT2	O60909	ND	1	B4GALT1	P15291	2AH9	1	50.0%	62.7%	1049.5	ND	ND
	B4GALT2	O60909	ND	1	LALBA	P00709	3B0O	2	4.0%	7.0%	7.5	ND	ND
	B4GALT1	P15291	2AH9	1	LALBA	P00709	3B0O	2	6.8%	12.0%	9.0	0.23387	5.70
2.4.2.31	SIRT6	Q8N6T7	3K35	1	ART1	P52961	ND	2	15.4%	23.7%	30.5	ND	ND
	SIRT6	Q8N6T7	3K35	1	ART3	Q13508	ND	2	18.6%	27.2%	43.5	ND	ND
	SIRT6	Q8N6T7	3K35	1	ART4	Q93070	ND	2	2.1%	3.2%	17.5	ND	ND
	SIRT6	Q8N6T7	3K35	1	ART5	Q96L15	ND	2	4.3%	5.7%	15.5	ND	ND
	ART1	P52961	ND	2	ART3	Q13508	ND	2	21.7%	33.0%	263.5	ND	ND
	ART1	P52961	ND	2	ART4	Q93070	ND	2	29.3%	42.7%	377.0	ND	ND
	ART1	P52961	ND	2	ART5	Q96L15	ND	2	34.8%	46.6%	447.5	ND	ND
	ART3	Q13508	ND	2	ART4	Q93070	ND	2	18.8%	30.5%	221.0	ND	ND
	ART3	Q13508	ND	2	ART5	Q96L15	ND	2	25.6%	35.1%	391.5	ND	ND
	ART4	Q93070	ND	2	ART5	Q96L15	ND	2	28.7%	43.9%	321.0	ND	ND
2.7.1.67	PI4KA	P42356	ND	1	PI4KB	Q9UBF8	4WAE	1	10.8%	17.1%	527.5	ND	ND
	PI4KA	P42356	ND	1	PI4K2A	Q9BTU6	4HND	2	3.8%	6.4%	46.0	ND	ND
	PI4KA	P42356	ND	1	PI4K2B	Q8TCG2	4WTV	2	4.7%	8.0%	28.5	ND	ND
	PI4KB	Q9UBF8	4WAE	1	PI4K2A	Q9BTU6	4HND	2	9.8%	16.4%	45.5	0.48577	4.85
	PI4KB	Q9UBF8	4WAE	1	PI4K2B	Q8TCG2	4WTV	2	9.9%	16.5%	51.0	0.31144	4.58
	PI4K2A	Q9BTU6	4HND	2	PI4K2B	Q8TCG2	4WTV	2	57.7%	69.5%	1472.5	0.46188	1.74
2.7.4.21	PPIP5K2	O43314	3T9A	1	PPIP5K1	Q6PFW1	ND	1	56.2%	64.5%	4170.5	ND	ND
	PPIP5K2	O43314	3T9A	1	IP6K1	Q92551	ND	2	8.0%	12.8%	50.0	ND	ND
	PPIP5K2	O43314	3T9A	1	IP6K3	Q96PC2	ND	2	7.1%	11.2%	49.5	ND	ND
	PPIP5K2	O43314	3T9A	1	IP6K2	Q9UHH9	ND	2	7.0%	11.3%	47.5	ND	ND
	PPIP5K1	Q6PFW1	ND	1	IP6K1	Q92551	ND	2	6.7%	10.5%	96.5	ND	ND
	PPIP5K1	Q6PFW1	ND	1	IP6K3	Q96PC2	ND	2	5.8%	10.2%	42.5	ND	ND
	PPIP5K1	Q6PFW1	ND	1	IP6K2	Q9UHH9	ND	2	6.1%	9.6%	43.5	ND	ND
	IP6K1	Q92551	ND	2	IP6K3	Q96PC2	ND	2	47.6%	61.4%	1072.0	ND	ND
	IP6K1	Q92551	ND	2	IP6K2	Q9UHH9	ND	2	46.3%	62.2%	1019.0	ND	ND
	IP6K3	Q96PC2	ND	2	IP6K2	Q9UHH9	ND	2	44.7%	58.7%	911.0	ND	ND

Tabela 4.1

EC	Gene[1]	Uniprot[1]	PDB[1]	Cluster[1]	Gene[2]	Uniprot[2]	PDB[2]	Cluster[2]	Identidade	Similaridade	*Score*	TM-Score	RMSD
	AADAC	P22760	ND	3	CEL	B4DSX9	ND	3	10.3%	17.7%	100.5	ND	ND
	AADAC	P22760	ND	3	LIPC	P11150	ND	4	17.1%	27.6%	26.0	ND	ND
	AADAC	P22760	ND	3	PNLIP	P16233	1LPB	4	13.6%	24.5%	18.0	ND	ND
	AADAC	P22760	ND	3	PNLIPRP1	P54315	2PPL	4	11.5%	20.7%	25.5	ND	ND
	AADAC	P22760	ND	3	PNLIPRP3	Q17RR3	ND	4	13.5%	22.3%	36.0	ND	ND
	AADAC	P22760	ND	3	LIPG	Q9Y5X9	ND	4	11.7%	22.4%	15.0	ND	ND
	AADAC	P22760	ND	3	PNLIPRP2	P54317	2OXE	4	14.6%	26.6%	36.0	ND	ND
	CEL	B4DSX9	ND	3	LIPC	P11150	ND	4	10.3%	15.8%	41.0	ND	ND
	CEL	B4DSX9	ND	3	PNLIP	P16233	1LPB	4	11.8%	18.7%	34.0	ND	ND
	CEL	B4DSX9	ND	3	PNLIPRP1	P54315	2PPL	4	5.6%	8.6%	29.0	ND	ND
	CEL	B4DSX9	ND	3	PNLIPRP3	Q17RR3	ND	4	7.4%	12.7%	31.5	ND	ND
	CEL	B4DSX9	ND	3	LIPG	Q9Y5X9	ND	4	11.5%	19.2%	42.5	ND	ND
	CEL	B4DSX9	ND	3	PNLIPRP2	P54317	2OXE	4	4.8%	8.5%	21.5	ND	ND
	AADAC	P22760	ND	3	LIPF	P07098	1HLG	9	18.2%	31.1%	57.0	ND	ND
	CEL	B4DSX9	ND	3	LIPF	P07098	1HLG	9	13.5%	22.3%	32.0	ND	ND
	AADAC	P22760	ND	3	PNPLA3	Q9NST1	ND	10	3.4%	5.1%	36.0	ND	ND
	CEL	B4DSX9	ND	3	PNPLA3	Q9NST1	ND	10	17.8%	26.5%	42.5	ND	ND
	LIPC	P11150	ND	4	PNLIP	P16233	1LPB	4	28.4%	42.1%	503.0	ND	ND
3.1.1.3	LIPC	P11150	ND	4	PNLIPRP1	P54315	2PPL	4	29.0%	42.3%	506.0	ND	ND
	LIPC	P11150	ND	4	PNLIPRP3	Q17RR3	ND	4	29.5%	43.9%	536.0	ND	ND
	LIPC	P11150	ND	4	LIPG	Q9Y5X9	ND	4	41.3%	61.0%	1059.5	ND	ND
	LIPC	P11150	ND	4	PNLIPRP2	P54317	2OXE	4	27.1%	43.0%	473.5	ND	ND
	PNLIP	P16233	1LPB	4	PNLIPRP1	P54315	2PPL	4	67.3%	80.6%	1750.0	0.93392	1.76
	PNLIP	P16233	1LPB	4	PNLIPRP3	Q17RR3	ND	4	47.3%	63.7%	1113.5	ND	ND
	PNLIP	P16233	1LPB	4	LIPG	Q9Y5X9	ND	4	30.5%	42.2%	556.5	ND	ND
	PNLIP	P16233	1LPB	4	PNLIPRP2	P54317	2OXE	4	64.0%	79.5%	1676.0	0.94537	1.37
	PNLIPRP1	P54315	2PPL	4	PNLIPRP3	Q17RR3	ND	4	48.4%	64.3%	1158.0	ND	ND
	PNLIPRP1	P54315	2PPL	4	LIPG	Q9Y5X9	ND	4	28.5%	42.5%	543.5	ND	ND
	PNLIPRP1	P54315	2PPL	4	PNLIPRP2	P54317	2OXE	4	62.7%	77.0%	1655.0	0.92898	1.81
	PNLIPRP3	Q17RR3	ND	4	LIPG	Q9Y5X9	ND	4	29.3%	44.2%	519.0	ND	ND
	PNLIPRP3	Q17RR3	ND	4	PNLIPRP2	P54317	2OXE	4	47.8%	62.2%	1156.5	ND	ND
	LIPG	Q9Y5X9	ND	4	PNLIPRP2	P54317	2OXE	4	28.6%	44.9%	536.5	ND	ND
	LIPC	P11150	ND	4	LIPF	P07098	1HLG	9	15.9%	28.2%	41.0	ND	ND
	PNLIP	P16233	1LPB	4	LIPF	P07098	1HLG	9	17.2%	26.2%	65.5	0.39307	4.45
	PNLIPRP1	P54315	2PPL	4	LIPF	P07098	1HLG	9	16.8%	29.2%	37.0	0.38582	5.08
	PNLIPRP3	Q17RR3	ND	4	LIPF	P07098	1HLG	9	16.5%	25.8%	41.5	ND	ND
	LIPG	Q9Y5X9	ND	4	LIPF	P07098	1HLG	9	12.3%	22.0%	26.0	ND	ND
	PNLIPRP2	P54317	2OXE	4	LIPF	P07098	1HLG	9	6.1%	9.8%	26.0	0.40059	5.14
	LIPC	P11150	ND	4	PNPLA3	Q9NST1	ND	10	6.9%	11.9%	48.5	ND	ND
	PNLIP	P16233	1LPB	4	PNPLA3	Q9NST1	ND	10	11.6%	18.8%	31.0	ND	ND
	PNLIPRP1	P54315	2PPL	4	PNPLA3	Q9NST1	ND	10	14.3%	21.2%	54.0	ND	ND

Tabela 4.1

EC	Gene[1]	Uniprot[1]	PDB[1]	Cluster[1]	Gene[2]	Uniprot[2]	PDB[2]	Cluster[2]	Identidade	Similaridade	*Score*	TM-Score	RMSD
	PNLIPRP3	Q17RR3	ND	4	PNPLA3	Q9NST1	ND	10	9.2%	16.5%	34.0	ND	ND
	LIPG	Q9Y5X9	ND	4	PNPLA3	Q9NST1	ND	10	7.9%	12.3%	30.0	ND	ND
	PNLIPRP2	P54317	2OXE	4	PNPLA3	Q9NST1	ND	10	13.1%	20.8%	38.5	ND	ND
	LIPF	P07098	1HLG	9	PNPLA3	Q9NST1	ND	10	2.5%	3.5%	7.0	ND	ND
	PTRH2	Q9Y3E5	1Q7S	1	PTRH1	Q86Y79	ND	2	17.6%	27.5%	28.0	ND	ND
3.1.1.29	PTRH2	Q9Y3E5	1Q7S	1	ICT1	Q14197	ND	3	10.2%	14.9%	18.0	ND	ND
	PTRH1	Q86Y79	ND	2	ICT1	Q14197	ND	3	13.3%	21.8%	15.5	ND	ND
	ACOT2	P49753	3HLK	1	BAAT	Q14032	ND	1	38.0%	51.0%	873.5	ND	ND
	ACOT2	P49753	3HLK	1	ACOT1	Q86TX2	ND	1	86.1%	86.5%	2217.0	ND	ND
	ACOT2	P49753	3HLK	1	ACOT4	Q8N9L9	3K2I	1	61.1%	70.6%	1601.0	0.95168	1.20
	BAAT	Q14032	ND	1	ACOT1	Q86TX2	ND	1	42.9%	56.9%	868.5	ND	ND
	BAAT	Q14032	ND	1	ACOT4	Q8N9L9	3K2I	1	43.1%	56.7%	841.0	ND	ND
3.1.2.2	ACOT1	Q86TX2	ND	1	ACOT4	Q8N9L9	3K2I	1	70.3%	81.0%	1603.0	ND	ND
	ACOT2	P49753	3HLK	1	ACOT7	O00154	2QQ2	2	2.8%	4.6%	27.5	0.19192	4.25
	BAAT	Q14032	ND	1	ACOT7	O00154	2QQ2	2	0.4%	0.8%	9.0	ND	ND
	ACOT1	Q86TX2	ND	1	ACOT7	O00154	2QQ2	2	13.4%	21.5%	18.5	ND	ND
	ACOT4	Q8N9L9	3K2I	1	ACOT7	O00154	2QQ2	2	2.0%	2.5%	13.5	0.23303	5.39
	ACP5	P13686	1WAR	1	ACP2	P11117	ND	2	4.2%	7.9%	23.0	ND	ND
	ACP5	P13686	1WAR	1	ACPP	P15309	1CVI	2	16.6%	28.3%	22.5	0.38800	5.44
	ACP5	P13686	1WAR	1	ACPT	Q9BZG2	ND	2	17.4%	24.4%	43.0	ND	ND
	ACP5	P13686	1WAR	1	ACP6	Q9NPH0	4JOB	2	15.5%	24.3%	19.5	0.36502	5.59
	ACP5	P13686	1WAR	1	ACP1	P24666	5PNT	5	10.1%	17.0%	24.5	0.34702	5.16
	ACP2	P11117	ND	2	ACPP	P15309	1CVI	2	43.6%	58.4%	976.5	ND	ND
	ACP2	P11117	ND	2	ACPT	Q9BZG2	ND	2	43.0%	57.2%	842.5	ND	ND
3.1.3.2	ACP2	P11117	ND	2	ACP6	Q9NPH0	4JOB	2	21.3%	33.9%	269.5	ND	ND
	ACPP	P15309	1CVI	2	ACPT	Q9BZG2	ND	2	36.8%	50.2%	770.0	ND	ND
	ACPP	P15309	1CVI	2	ACP6	Q9NPH0	4JOB	2	26.1%	41.5%	319.5	0.80111	2.77
	ACPT	Q9BZG2	ND	2	ACP6	Q9NPH0	4JOB	2	24.4%	35.3%	289.5	ND	ND
	ACP2	P11117	ND	2	ACP1	P24666	5PNT	5	8.5%	13.9%	28.5	ND	ND
	ACPP	P15309	1CVI	2	ACP1	P24666	5PNT	5	10.1%	18.9%	19.0	0.26430	5.89
	ACPT	Q9BZG2	ND	2	ACP1	P24666	5PNT	5	6.0%	12.9%	13.0	ND	ND
	ACP6	Q9NPH0	4JOB	2	ACP1	P24666	5PNT	5	2.2%	4.1%	16.5	0.24707	4.77
	NT5C1B	Q96P26	ND	2	NT5C1A	Q9BXI3	ND	2	35.4%	43.0%	1145.0	ND	ND
	NT5C1B	Q96P26	ND	2	NT5E	P21589	4H2G	3	8.5%	13.3%	26.0	ND	ND
	NT5C1A	Q9BXI3	ND	2	NT5E	P21589	4H2G	3	10.8%	20.1%	32.5	ND	ND
	NT5C1B	Q96P26	ND	2	NT5C	Q8TCD5	4L57	5	7.3%	13.2%	29.5	ND	ND
3.1.3.5	NT5C1B	Q96P26	ND	2	NT5M	Q9NPB1	4MUM	5	6.0%	9.3%	48.0	ND	ND
	NT5C1A	Q9BXI3	ND	2	NT5C	Q8TCD5	4L57	5	9.7%	15.1%	29.0	ND	ND
	NT5C1A	Q9BXI3	ND	2	NT5M	Q9NPB1	4MUM	5	8.5%	14.0%	26.5	ND	ND
	NT5C1B	Q96P26	ND	2	NT5C3A	Q9H0P0	2CN1	7	8.6%	16.6%	35.5	ND	ND

Tabela 4.1

EC	Gene[1]	Uniprot[1]	PDB[1]	Cluster[1]	Gene[2]	Uniprot[2]	PDB[2]	Cluster[2]	Identidade	Similaridade	Score	TM-Score	RMSD
	NT5C1A	Q9BXI3	ND	2	NT5C3A	Q9H0P0	2CN1	7	9.1%	17.1%	23.5	ND	ND
	NT5C1B	Q96P26	ND	2	NT5C2	P49902	2XCW	9	8.6%	13.9%	42.5	ND	ND
	NT5C1A	Q9BXI3	ND	2	NT5C2	P49902	2XCW	9	8.2%	14.5%	20.0	ND	ND
	NT5E	P21589	4H2G	3	NT5C	Q8TCD5	4L57	5	7.7%	13.4%	7.0	0.26104	5.86
	NT5E	P21589	4H2G	3	NT5M	Q9NPB1	4MUM	5	2.0%	3.2%	12.0	0.26951	5.69
	NT5E	P21589	4H2G	3	NT5C3A	Q9H0P0	2CN1	7	8.9%	15.9%	30.5	0.26314	5.72
	NT5E	P21589	4H2G	3	NT5C2	P49902	2XCW	9	15.4%	26.8%	36.0	0.27774	7.25
	NT5C	Q8TCD5	4L57	5	NT5M	Q9NPB1	4MUM	5	51.3%	64.7%	660.0	0.96844	0.70
	NT5C	Q8TCD5	4L57	5	NT5C3A	Q9H0P0	2CN1	7	6.7%	9.9%	27.5	0.50279	4.85
	NT5M	Q9NPB1	4MUM	5	NT5C3A	Q9H0P0	2CN1	7	14.1%	26.9%	40.0	0.50929	4.86
	NT5C	Q8TCD5	4L57	5	NT5C2	P49902	2XCW	9	7.5%	11.8%	26.0	0.44003	3.79
	NT5M	Q9NPB1	4MUM	5	NT5C2	P49902	2XCW	9	4.6%	9.1%	23.5	0.44594	3.87
	NT5C3A	Q9H0P0	2CN1	7	NT5C2	P49902	2XCW	9	8.5%	14.7%	46.0	0.45164	4.71
	SMPD2	O60906	ND	1	SMPD3	Q9NY59	ND	1	10.4%	15.2%	87.5	ND	ND
	SMPD2	O60906	ND	1	SMPD1	P17405	5I81	2	5.5%	10.0%	35.5	ND	ND
	SMPD3	Q9NY59	ND	1	SMPD1	P17405	5I81	2	2.6%	3.6%	51.0	ND	ND
	SMPD2	O60906	ND	1	SMPD4	Q9NXE4	ND	3	8.2%	13.6%	62.0	ND	ND
3.1.4.12	SMPD3	Q9NY59	ND	1	SMPD4	Q9NXE4	ND	3	12.7%	20.2%	51.5	ND	ND
	SMPD2	O60906	ND	1	ENPP7	Q6UWV6	5UDY	4	7.0%	12.8%	27.5	ND	ND
	SMPD3	Q9NY59	ND	1	ENPP7	Q6UWV6	5UDY	4	13.3%	19.2%	29.5	ND	ND
	SMPD1	P17405	5I81	2	SMPD4	Q9NXE4	ND	3	6.3%	10.7%	39.0	ND	ND
	SMPD1	P17405	5I81	2	ENPP7	Q6UWV6	5UDY	4	16.1%	25.0%	28.5	0.28698	6.61
	SMPD4	Q9NXE4	ND	3	ENPP7	Q6UWV6	5UDY	4	5.6%	8.7%	49.5	ND	ND

Tabela 4.1

EC	Gene[1]	Uniprot[1]	PDB[1]	Cluster[1]	Gene[2]	Uniprot[2]	PDB[2]	Cluster[2]	Identidade	Similaridade	Score	TM-Score	RMSD
	NTHL1	P78549	ND	1	OGG1	O15527	1KO9	1	19.0%	28.4%	85.5	ND	ND
	NTHL1	P78549	ND	1	NEIL2	Q969S2	1VZP	1	14.6%	25.7%	29.0	ND	ND
	NTHL1	P78549	ND	1	NEIL1	Q96FI4	1TDH	1	6.7%	9.8%	40.0	ND	ND
	OGG1	O15527	1KO9	1	NEIL2	Q969S2	1VZP	1	2.9%	4.3%	24.0	0.27899	4.84
	OGG1	O15527	1KO9	1	NEIL1	Q96FI4	1TDH	1	14.3%	20.2%	42.5	0.28166	6.65
	NTHL1	P78549	ND	1	APEX1	P27695	2O3H	1	15.6%	23.0%	41.0	ND	ND
	NTHL1	P78549	ND	1	APEX2	Q9UBZ4	ND	1	4.7%	6.6%	14.5	ND	ND
	OGG1	O15527	1KO9	1	APEX1	P27695	2O3H	1	16.5%	26.7%	23.5	0.27998	6.74
	OGG1	O15527	1KO9	1	APEX2	Q9UBZ4	ND	1	9.4%	14.6%	42.5	ND	ND
	NTHL1	P78549	ND	1	APLF	Q8IW19	2KUO	6	3.7%	6.1%	24.5	**ND**	**ND**
4.2.99.18	OGG1	O15527	1KO9	1	APLF	Q8IW19	2KUO	6	6.0%	8.7%	20.5	0.14656	6.94
	NEIL2	Q969S2	1VZP	1	NEIL1	Q96FI4	1TDH	1	18.8%	24.9%	143.5	0.50213	2.44
	NEIL2	Q969S2	1VZP	1	APEX1	P27695	2O3H	1	13.2%	22.2%	19.5	0.28098	4.80
	NEIL2	Q969S2	1VZP	1	APEX2	Q9UBZ4	ND	1	6.6%	10.5%	42.0	ND	ND
	NEIL1	Q96FI4	1TDH	1	APEX1	P27695	2O3H	1	5.2%	7.3%	44.0	0.25323	6.80
	NEIL1	Q96FI4	1TDH	1	APEX2	Q9UBZ4	ND	1	14.6%	20.8%	41.0	ND	ND
	NEIL2	Q969S2	1VZP	1	APLF	Q8IW19	2KUO	6	3.9%	7.9%	24.5	0.09438	5.32
	NEIL1	Q96FI4	1TDH	1	APLF	Q8IW19	2KUO	6	10.7%	18.3%	46.5	0.15631	6.56
	APEX1	P27695	2O3H	1	APEX2	Q9UBZ4	ND	1	14.9%	22.0%	264.0	ND	ND
	APEX1	P27695	2O3H	1	APLF	Q8IW19	2KUO	6	6.7%	12.8%	47.5	0.16022	6.62
	APEX2	Q9UBZ4	ND	1	APLF	Q8IW19	2KUO	6	12.3%	20.5%	44.5	ND	ND
5.3.99.2	PTGDS	P41222	2WWP	1	HPGDS	O60760	1IYI	2	13.7%	21.4%	9.5	0.28983	5.76
5.3.99.3	PTGES2	Q9H7Z7	ND	1	PTGES	O14684	4AL0	2	0.6%	0.6%	18.0	ND	ND

*** O modelo 3D da desidrogenase humana (UniProt Q9UQ10) foi obtido utilizando a estrutura cristalina de uma desidrogenase de *Macaca fascicularis* (PDB 2O48) por modelagem comparativa. [1] informações sobre a primeira proteína usada na comparação par a par. [2] informações sobre a segunda enzima usada na comparação par a par**

Em resumo, foi possível inferir convergência funcional em todas as 15 atividades enzimáticas com base em informações de domínios e superfamílias, além do alinhamento estrutural entre formas alternativas em nove dessas 15 atividades enzimáticas (ECs 1.3.1.20, 1.15.1.1, 2.4.1.22, 2.7.1.67, 3.1.1.3, 3.1.2.2, 3.1.3.2, 3.1.3.5, 4.2.99.18 e 5.3.99.2). Como mostrado na Figura 4.2, com exceção dos genes PTGES e PTGES2, codificadores de enzimas da atividade 5.3.99.3, no cromossomo 9, os genes codificadores de enzimas análogas intragenômicas aparentam estar aleatoriamente distribuídos, dispersos ao longo de todo o genoma humano, sendo observados em 21 dos 24 cromossomos nucleares (20 autossômicos e 1 cromossomo sexual). Para os genes codificadores de formas alternativas, assim como genes codificadores de formas enzimáticas homólogas, foram mapeadas as localizações cromossômicas e então construído um diagrama circular, como representado na Figura 4.2. De maneira similar, as distâncias entre os genes codificadores de análogos intragenômicos e entre enzimas homólogas exibem um padrão disperso ao longo do genoma.

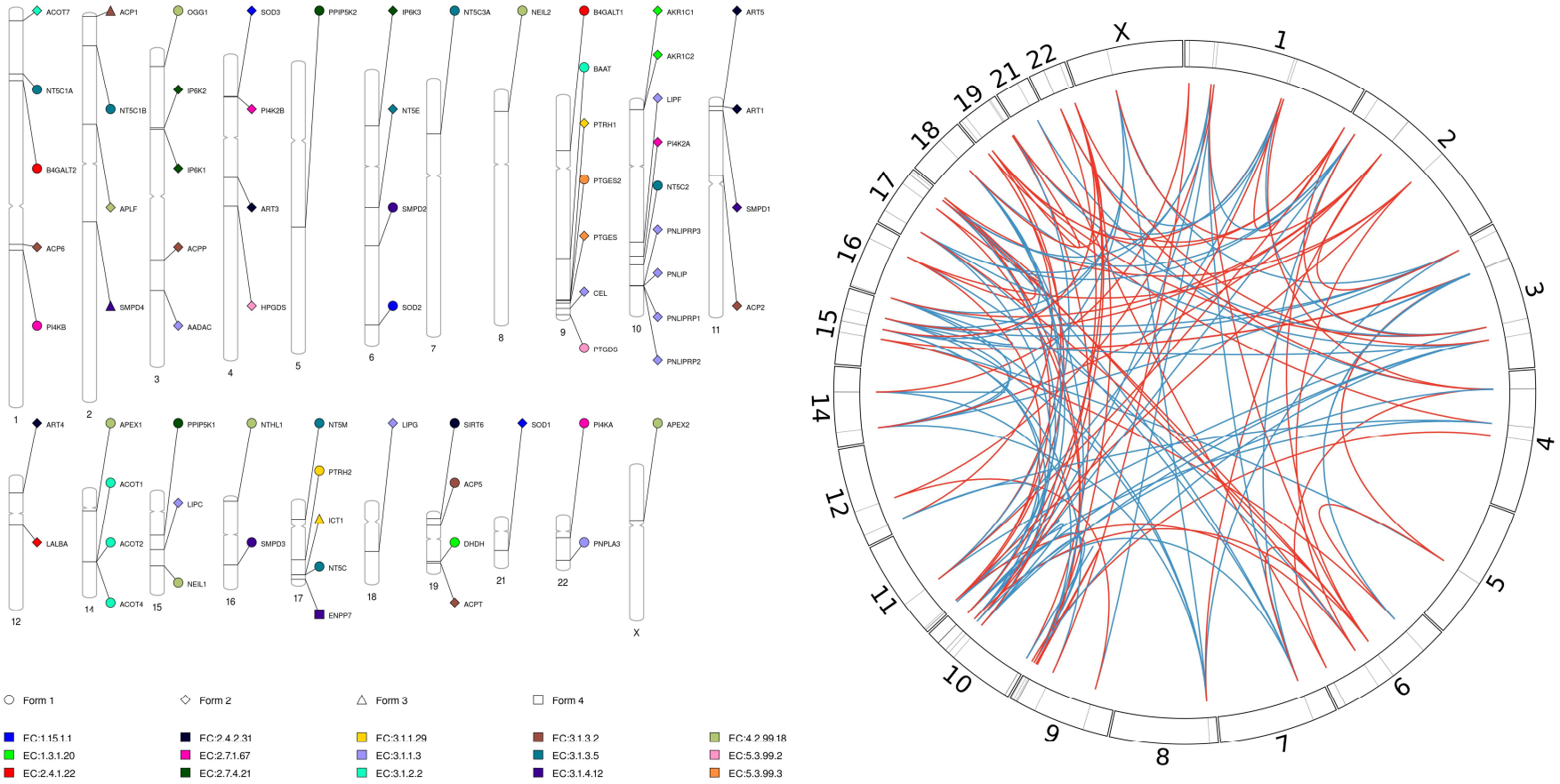


Figura 4.2 (Esquerda) Diagrama representando a localização dos genes codificadores de enzimas análogas intragenômicas ao longo do cromossomo humano. As atividades enzimáticas com evidência de analogia intragenômica encontradas estão representadas por cores distintas. Genes codificando distintas formas enzimáticas estão representados por diferentes símbolos. (Direita) Diagrama circular apresentando as distâncias entre os genes codificadores de formas alternativas (pertencentes a grupos distintos formados pelo AnEnPi em um determinado EC) em linhas vermelhas e genes codificadores de formas enzimáticas homólogas (pertencentes ao mesmo grupo formado pelo AnEnPi em um determinado EC) através de linhas azuis. Os cromossomos humanos estão representados como segmentos contínuos em um círculo, no qual as linhas verticais pretas ao longo desses cromossomos correspondem às coordenadas cromossômicas dos 70 genes codificadores de enzimas análogas encontrados no conjunto de dados (+) *bona fide*. Linhas curtas (vermelhas e azuis) correspondem aos genes vizinhos em um cromossomo.

4.2 Nucleotidades, Lipases, Desidrogenases, Sintases e Dismutases

A literatura científica indica a existência de sete 5′-nucleotidases (EC 3.1.3.5) em humanos, que são hidrolases envolvidas na biossíntese de nucleosídeos e fosfato inorgânico a partir de nucleosídeos monofosfatados não cíclicos, codificadas pelos genes NT5E, NT5C1A, NT5C1B, NT5C, NT5C3A, NT5C2 e NT5M: uma enzima solúvel associada a membrana plasmática (NT5E), e seis enzimas com localização intracelular, tanto citossólica (NT5C1A, NT5C1B, NT5C, NT5C3A e NT5C2) como mitocondrial (NT5M) (85). Todos esses genes estão distribuídos em cromossomos distintos (1, 2, 6, 7, 10 e 17) (Figura 4.2), e as enzimas codificadas por eles foram atribuídas a 5 grupos do AnEnPi: (i) NT5E, (ii) NT5C1A e NT5C1B, (iii) NT5C e NT5M, (iv) NT5C3A e (v) NT5C2 (Anexo I). A enzima codificada pelo gene NT5E pertence à superfamília *5′-nucleotidase C-terminal domain e Metallo-dependent phosphatases*, enquanto a enzima codificada pelo gene NT5C, assim como as demais 5′-nucleotidases (NT5C3A, NT5C2, NT5M), são membros da superfamília *HAD-like*. As enzimas codificadas pelos genes NT5C1A e NT5C1B não possuem anotação de superfamília ou estrutura 3D, mas foram agrupadas juntas, mostrando considerável similaridade de sequência, indicando uma possível origem comum (Tabela 4.1). A enzima associada à membrana, NT5E, claramente se distingue das demais formas enzimáticas em todas as métricas utilizadas, uma vez que ela foi alocada em um grupo completamente separado pelo AnEnPi, exibindo valores marcadamente baixos de similaridade de sequência e estrutura quando comparada às demais 5′-nucleotidases, possuindo uma classificação de superfamília/enovelamento completamente diferente (como mencionado anteriormente), e composição/arquitetura de domínios distinta (Tabela 4.1 e Anexo I). Por outro lado, as enzimas citossólicas, membros da superfamília *HAD-like*, codificadas pelos genes NT5C, NT5C3A e NT5C2, assim como a forma mitocondrial, codificada pelo gene NT5M, foram atribuídas a três grupos

separados pelo AnEnPi; as enzimas codificadas pelos genes NT5C e NT5M, pertencentes ao mesmo grupo, exibem alta similaridade estrutural e de sequência entre elas, assim como a mesma composição/arquitetura de domínios, enquanto uma tendência oposta é observada quando enzimas representantes da superfamília *HAD-like* de diferentes grupos são comparadas (NT5C ou NT5M contra NT5C3A ou NT5C2, e NT5C3A contra NT5C2): similaridade de sequência e estrutural muito baixas e composição/arquitetura de domínios não relacionadas (Tabela 4.1 e Anexo I). Curiosamente, Crisp e colaboradores (2015) mostraram considerável evidência de que os genes NT5C e NT5M tenham sido transferidos horizontalmente para a linhagem humana (possivelmente vindo de genomas bacterianos), desta forma, contribuindo para a diversificação bioquímica de 5′-nucleotidases ao longo da evolução animal (93). Apesar da diversidade de localizações subcelulares, possível origem evolutiva, sequência de aminoácido, composição/arquitetura de enovelamento e domínio, essas enzimas usam 5′-nucleotídeos de várias fontes, exibindo diferenças significativas na amplitude de substratos (parcialmente sobrepostas), assim como na especificidade de substratos (94). Portanto, é razoável pensar na possibilidade de que essas enzimas cumpram diferentes papéis biológicos durante a regulação de diversos processos fisiológicos.

As oxirredutases *Trans-1,2-Dihydrobenzene-1,2-Diol Dehydrogenase* (EC 1.3.1.20) compreendem enzimas codificadas pelos genes DHDH, AKR1C1 e AKR1C2. Em nossas análises, o AnEnPi atribuiu a enzima codificada pelo gene DHDH a um grupo separado, enquanto as enzimas restantes (codificadas pelos genes AKR1C1 e AKR1C2) foram reunidas em um grupo diferente. As formas análogas preditas puderam ser abalizadas com base em composição/arquitetura de domínios, anotação de superfamília, assim como estrutura 3D (Tabela 4.1 e Anexos I e II). Vale ressaltar que o gene DHDH está localizado no cromossomo 19, enquanto os demais

são vizinhos, co-localizados no cromossomo 10, com os seus produtos apresentando sequências de aminoácidos praticamente idênticas (97,8% de identidade entre todas as sequências), portanto, reforçando a evidência de origem comum (possivelmente duplicação) para os genes AKR1C1 e AKR1C2 (Figura 4.2). A baixa similaridade de sequência e estrutural entre a enzima DHDH e membros da família *aldo-keto redutase* (*e.g.*, enzimas AKR1C1 e AKR1C2) já havia sido descrita, assim como diferenças no uso de substratos (88, 89). A enzima DHDH atua sobre *(–)-[1R,2R]-dihydrodiols*, enquanto *aldo-keto reductases* oxidam *(+)-[1S,2S]-dihydrodiols* (89). Além disso, membros da família *aldo-keto redutase* utilizam esteroides sintéticos como substrato (95). É sabido que enzimas homólogas podem também apresentar diferentes especificidades de substratos, mas, nesse caso, a diferença estabelecida de substrato se correlaciona com a presumida origem evolutiva distinta, mesmo na ausência de informações adicionais que possam indicar outras implicações possíveis em papéis biológicos distintos.

Representantes da classe enzimática isomerase, *prostaglandin D2 synthase* e *hematopoietic prostaglandin D synthase* (codificadas pelos genes PTGDS, localizado no cromossomo 9 e HPGDS, localizado no cromossomo 4, respectivamente) (EC 5.3.99.2), ambas as enzimas regulam a síntese de prostaglandina D2, atuando na sinalização e processos inflamatórios (96–98). O *pipeline* AnEnPi atribuiu as enzimas codificadas pelos genes PTGDS e HPGDS a grupos diferentes, e análises subsequentes revelaram que essas enzimas também não são relacionadas entre si em termos de composição de domínios, classificação de superfamília, assim como sequência de aminoácido e estrutura 3D (Tabela 4.1 e Anexo I e II), corroborando evidências prévias sobre convergência funcional nessa atividade enzimática (96,97). De acordo com esse achado, uma busca na literatura científica revela inúmeras características que poderiam distinguir papéis distintos para estas enzimas, tais como

(i) a presença de peptídeo sinal e sítios de N-glicosilação apenas na enzima PTGDS (92); (ii) distintas localizações teciduais, bem como inibidores e ativadores, que podem estar relacionados com diferentes mecanismos de ação (96); (iii) a enzima PTGDS é secretada e é preferencialmente expressa no cérebro, estando também envolvida com a regulação do sono, adipogênese, resposta alérgica e inflamatória (98–100); (iv) a enzima HPGDS está presente em células do sistema imune (101).

Outra atividade enzimática encontrada em todos os seres vivos é a (oxidoredutase) superoxido dismutase (SOD) (EC 1.15.1.1); as enzimas SOD catalisam a conversão de radicais superóxido ($O2^{-}$) em peróxido de hidrogênio (H_2O_2) ou oxigênio molecular (O_2), protegendo as células, tecidos e órgãos do estresse oxidativo. Humanos e demais mamíferos expressam três formas de SOD: SOD1, enzima citoplasmática dependente de cobre/zinco (codificada pelo gene SOD1 no cromossomo 21); SOD2, enzima mitocondrial dependente de manganês (codificada pelo gene SOD2 no cromossomo 6); e SOD3, enzima extracelular dependente de cobre/zinco (codificada pelo gene SOD3 no cromossomo 4) (102). Através da nossa predição computacional de análogos funcionais, SOD1 e SOD3 foram reunidas em um mesmo grupo gerado pelo AnEnPi, enquanto SOD2 foi atribuída a um grupo separado, indicando um possível evento de origem *de novo*. Em análises subsequentes de composição de domínio, classificação de superfamília, sequência de aminoácido e estrutura 3D (Tabela 4.1 e Anexo I e II) foi possível confirmar que essas enzimas de fato não são relacionadas entre si, corroborando evidências sobre convergência funcional na atividade enzimática SOD, previamente descritas (34). Em um estudo recente, Garcia e colaboradores (2017) demonstraram que as enzimas dependentes de manganês com atividade superoxido dismutase, SodA e SodM, não apenas coexistem no patógeno humano *Staphylococcus aureus*, mas também exibem distintos papéis biológicos, no qual apenas uma das formas alternativas, SodM, pode

promover resistência a antibióticos e ao sistema imune do hospedeiro. Os autores mostraram que SodA é estritamente dependente do manganês e é relevante para o combate ao estresse oxidativo, bem como para o desenvolvimento de doenças quando o manganês é abundante, enquanto SodM é cambialista, essencial em condições de baixa disponibilidade de manganês, mantendo a mesma atividade enzimática na presença de manganês ou ferro (103). Mesmo que este fenômeno só tenha sido demonstrado em bactérias até o momento, ele abre a oportunidade de explorá-lo em outras espécies procarióticas ou eucarióticas.

Membros da classe transferase, enzimas com atividade *1-phosphatidylinositol 4-kinase* (PI4Ks) (EC 2.7.1.67) participam do metabolismo de inositol fosfato e do sistema de sinalização do fosfatidilinositol, catalisando a fosforilação do fosfatidilinositol. O produto dessa reação é o fosfatidilinositol 4-fosfato, um precursor primário na síntese de fosfatidilinositol polifosfatos, moléculas envolvidas em vários processos biológicos, como transdução de sinal, tráfego membranar e reorganização do citoesqueleto (104). As PI4Ks de mamíferos são classificadas em dois tipos, II e III, com base em características físico-químicas, e a literatura destaca a existência de diferentes organizações de domínios entre PI4Ks do tipo II (genes PI4K2A e PI4K2B) e PI4Ks do tipo III (genes PI4KA e PI4KB), com PI4KA e PI4KB sendo mais similares entre si, e PI4KA tendo um domínio de ligação característico (105,106). Portanto, a divisão de PI4Ks humanas em dois grupos distintos do AnEnPi, refletindo possíveis casos de formas isofuncionais não-homólogas – sendo um desses grupos formado pelos produtos dos genes PI4KA e PI4KB, e o outro correspondendo às enzimas codificadas pelos genes PI4K2A e PI4K2B, representantes de PI4Ks de mamíferos tipo III e II, respectivamente –, assim como a classificação dessas enzimas em distintas classes de superfamílias (exceto para a enzima codificada pelo gene PI4K2B, a qual não possui classificação de superfamília) e a suas estruturas 3D não

relacionadas (Anexo I e II), reforçam resultados similares obtidos previamente em estudos sobre essa atividade enzimática (106).

Ao todo, as comparações do tipo todos contra todos, entre pares de sequências de cada uma das atividades enzimáticas contidas no conjunto de dados (+) *bona fide* corroboraram as predições computacionais obtidas pelo AnEnPi. Entretanto, foram encontrados, pelo menos, dois casos em que o método de agrupamento usado pelo AnEnPi pode ter "produzido" mais formas enzimáticas alternativas humanas do que o esperado: EC 3.1.3.5 (*5'-nucleotidase*), com cinco *clusters*, e EC 3.1.1.3 (*Triacylglycerol Lipase*), com quatro clusters. No EC 3.1.3.5, os genes codificadores de enzimas NT5C, NT5M, NT5C3A e NT5C2 (formas citoplasmáticas) compartilham a mesma superfamília (*HAD-like*), enquanto a enzima codificada pelo gene NT5E (forma membranar) é simultaneamente classificada em duas superfamílias: *5'-nucleotidase* (*syn. UDP-sugar hydrolase*), *C-terminal domain* e *Metallo-dependent phosphatases*. Similarmente, os produtos dos genes AADAC, CEL, LIPC, PNLIP, PNLIPRP1, PNLIPRP3, LIPG e LIPF, compreendendo o EC 3.1.1.3, estão todas atribuídas à superfamília *alpha/beta-Hydrolases*, enquanto o produto do gene PNPLA3 pertence a uma distinta superfamília (*FabD/lysophospholipase-like*). Outra evidência que dá suporte a essa colocação é que os genes PNLIPRP1, PNLIPRP2, PNLIPRP3, PNLIP e LIPF são todos vizinhos, localizados no cromossomo humano 10, e suas enzimas correspondentes compartilham similaridade consideravelmente elevada entre si do que com a enzima PNPLA3, possivelmente representando eventos de duplicação gênica (Figura 4.2 e Tabela 4.1).

Como esperado, todas as enzimas isofuncionais não-homólogas estudadas estão associadas a distintos grupos de ortólogos do KEGG (*KEGG Orthologous*; KOs) <https://www.kegg.jp/kegg/ko.html>, corroborando a origem evolutiva distinta para as formas alternativas preditas (Anexo I). Apenas seis KOs são compartilhados entre

duas ou mais sequências no nosso conjunto de dados (+) *bona fide*: K01081, agrupando seis das sete *5'-nucleotidases* (NT5C1B, NT5C1A, NT5C, NT5M, NT5C3A e NT5C2); K01046, incluindo duas das dez *triacylglycerol lipases* (LIPC e LIPG); K01068, contendo três das cinco *palmitoyl-CoA hydrolases* (ACOT2 ACOT1 e ACOT4); K07756 e K13024 contendo todas as cinco *inositol-hexakisphosphate kinases* (IP6K1, IP6K3, IP6K2 e PPIP5K2, PPIP5K1, respectivamente); e K13711, agrupando duas das quatro *1-phosphatidylinositol 4-kinases* (PI4K2A e PI4K2B).

Após uma extensa pesquisa de literatura científica, não conseguimos encontrar informações sobre os genes codificadores de formas análogas nas nove atividades enzimáticas restantes de nosso conjunto de dados (+) *bona fide*: EC 2.4.1.22 (*Lactose synthase*), EC 2.4.2.31 (*NAD+—protein-arginine ADP-ribosyltransferase*), EC 2.7.4.21 (*Inositol-hexakisphosphate kinase*), EC 3.1.1.29 (*Aminoacyl-tRNA hydrolase*), EC 3.1.1.3 (*Triacylglycerol lipase*), EC 3.1.2.2 (*Palmitoyl-CoA hydrolase*), EC 3.1.3.2 (*Acid phosphatase*), EC 3.1.4.12 (*Sphingomyelin phosphodiesterase*), EC 4.2.99.18 (*DNA-(apurinic or apyrimidinic site) lyase*), e EC 5.3.99.3 (*Prostaglandin-E synthase*).

4.3 Análise Comparativa do Perfil de Transcrição, Redes de interação com miRNAs e Localização Subcelular dos Genes Codificadores de Formas Enzimáticas Análogas Intragenômicas Humanas

Todos os 70 genes codificadores de enzimas análogas intragenômicas humanas estavam presentes nos dados disponibilizados na base de dados GTEx, antes de qualquer processamento dos dados. O tamanho das bibliotecas nestes dados variou de aproximadamente 8,7 milhões de leituras até 312 milhões, com tamanho médio de 54,63 milhões. Dos 56.238 elementos genômicos presentes no conjunto de dados, 14.908 (~26%) apresentaram valor de CPM superior a 1 em pelo menos 50% das amostras trabalhadas (4.277 amostras). Entretanto, conforme pode ser visto no

gráfico da Figura 4.3, os dados exibem uma grande dispersão, com coeficiente de variação biológica de 0.1835701. Vale destacar que, além de trechos codificadores de proteínas, os dados do GTEx apresentaram valores de contagens para outros elementos genômicos, como pseudogenes e miRNAs, o que pode talvez possa explicar a grande variação observada.

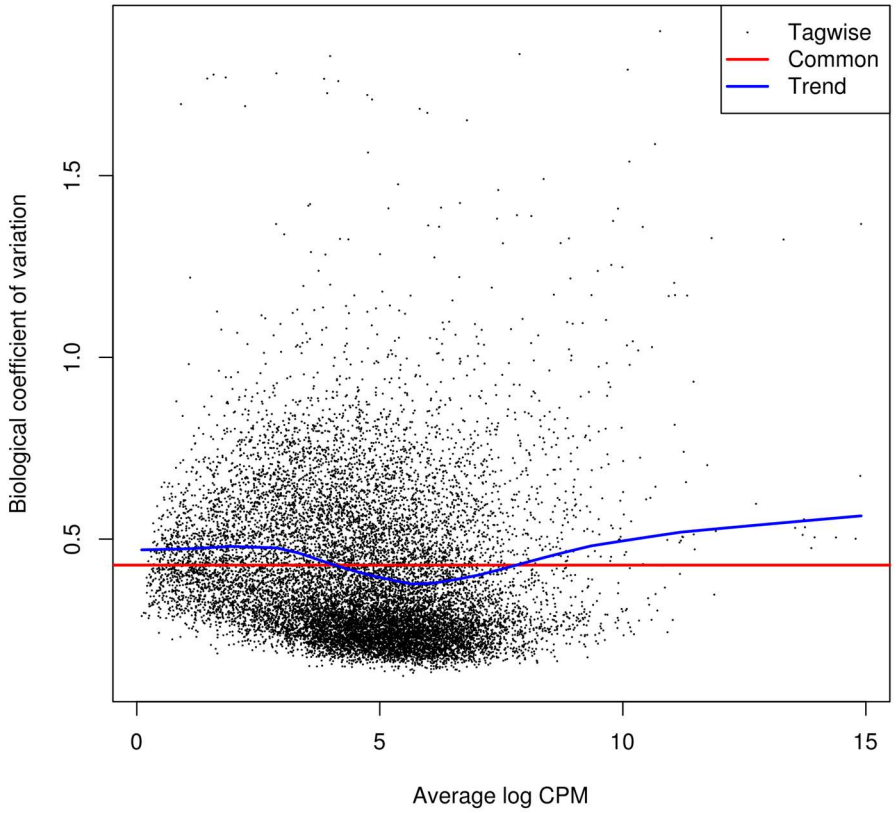


Figura 4.3 Coeficiente de variação biológica dos 14.908 elementos genômicos que apresentaram valor de CPM superior a 1 em pelo menos 50% das amostras trabalhadas (~26% do conjunto de dados inicial), oriundas do GTEx. A linha vermelha representa o valor de dispersão comum a todos os elementos; a linha azul representa o comportamento esperado; os pontos representam a dispersão de cada elemento genômico.

Após a remoção de possíveis artefatos do conjunto de dados (genes que apresentavam valor de CPM inferior a 1 em menos de 50% das amostras), restaram 42 genes codificadores de enzimas isofuncionais não-homólogas, representando 14 atividades enzimáticas distintas. No entanto, ao considerarmos somente os casos em que pelo menos dois genes codificadores de formas alternativas pertencentes à

mesma atividade enzimática, estavam representados neste conjunto de dados, o número final foi de 31 genes codificadores de formas análogas em 7 atividades enzimáticas (Tabela 4.2).

Uma das formas de se comparar o perfil de expessão de um grupo de genes é através da utilização de algoritmos de agrupamento, que permitem revelar genes com perfis de expressão similares. O método utilizado neste trabalho foi o agrupamento hierárquico aglomerativo, usando a fórmula (1-*Pearson Correlation*)/2 como medida de distância. Os motivos para usar tal métrica são os seguintes: matrizes de distância sempre contém valores positivos enquanto o coeficiente de Pearson pode ser negativo, e quanto menor for a distância, mais próximo (similar) os agrupamentos a serem combinados; o uso do coeficiente de Pearson em vez das técnicas padrão, como distância euclidiana, se aplica porque desejamos encontrar genes com perfis de expressão similares ao invés de capturar a magnitude da expressão (107). Dessa forma, o resultado deste algoritmo é um dendrograma que representa a correlação entre perfis de expressão (qualitativa).

Assim, com base nas médias dos valores normalizados por CPM, foram construídos agrupamentos hierárquicos para os genes codificadores de enzimas isofuncionais não-homólogas identificados em cada uma das atividades enzimáticas, permitindo reconhecer quais formas são co-expressas e quais possuem correlação negativa em nosso conjunto de dados. Além de disso, realizamos a construção de redes de interação com miRNAs e seus genes-alvos, buscando identificar miRNA humanos que possam estar envolvidos na modulação da expressão dos genes codificadores de enzimas análogas estudados, bem como reunimos informações sobre as localizações subcelulares das enzimas codificadas por esses genes, buscando caracterizar melhor as diferenças no perfil de expressão desses genes em distintos tecidos (Tabela 4.2 e Figura 4.4).

Tabela 4.2 Conjunto de genes codificadores de enzimas análogas intragenômicas humanas cujo perfil transcricional foi analisado neste trabalho. Cores distintas representam genes codificadores de distintas enzimas isofuncionais não-homólogas.

CLASSE	EC	ATIVIDADE	GENE	MIRNA	LOCALIZAÇÃO
Oxidorredutase	**1.15.1.1**	***superoxide dismutase***	**SOD2**	miR-330-3p	ND
			SOD1	ND	*Nucleus;Plasma membrane;Cytosol*
			SOD3	ND	ND
Transferase	**2.7.1.67**	***1-phosphatidylinositol 4-kinase***	**PI4KA**	miR-130a-3p; miR-130b-3p; miR-148a-3p; miR-148b-3p; miR-152-3p; miR-199a-5p; miR-199b-5p; miR-19a-3p; miR-19b-3p; miR-219a-5p; miR-301a-3p; miR-301b-3p; miR-454-3p	*Nucleoplasm;Plasma membrane*
			PI4KB	miR-103a-3p; miR-107; miR-140-3p.1; miR-140-3p.2; miR-15a-5p; miR-15b-5p; miR-16-5p; miR-195-5p; miR-200b-3p; miR-200c-3p; miR-369-3p; miR-424-5p; miR-429; miR-497-5p; miR-873-5p.1	*Golgi apparatus*
			PI4K2A	miR-140-3p.1; miR-15a-5p; miR-15b-5p; miR-16-5p; miR-181a-5p; miR-181b-5p; miR-181c-5p; miR-181d-5p; miR-182-5p; miR-195-5p; miR-204-5p; miR-211-5p; miR-218-5p; miR-323a-3p; miR-424-5p; miR-497-5p; miR-9-5p	*Plasma membrane*
			PI4K2B	let-7a-5p; let-7b-5p; let-7c-5p; let-7d-5p; let-7e-5p; let-7f-5p; let-7g-5p; let-7i-5p; miR-124-3p.1; miR-124-3p.2; miR-125a-5p; miR-125b-5p; miR-133a-3p.1; miR-145-5p; miR-15a-5p; miR-15b-5p; miR-16-5p; miR-181a-5p; miR-181b-5p; miR-181c-5p; miR-181d-5p; miR-195-5p; miR-200b-3p; miR-200c-3p; miR-30a-5p; miR-30b-5p; miR-30c-5p; miR-30d-5p; miR-30e-5p; miR-338-3p; miR-424-5p; miR-429; miR-497-5p; miR-503-5p; miR-506-3p; miR-98-5p	*Cytosol*

Tabela 4.2

CLASSE	EC	ATIVIDADE	GENE	MIRNA	LOCALIZAÇÃO
Hidrolase	3.1.2.2	*palmitoyl-CoA hydrolase*	ACOT2	ND	*Mitochondria*
			ACOT1	ND	*Mitochondria*
			ACOT4	ND	ND
			ACOT7	miR-141-3p; miR-200a-3p; miR-9-5p	*Nucleoplasm;Cytosol*
	3.1.3.2	*acid phosphatase*	ACP5	ND	*Cytosol*
			ACP2	ND	ND
			ACP6	miR-378a-3p; miR-378b; miR-378c; miR-378d; miR-378e; miR-378f; miR-378h; miR-378i	ND
			ACP1	miR-122-5p; miR-129-5p; miR-19a-3p; miR-19b-3p; miR-379-3p; miR-411-3p; miR-494-3p	*Nucleus;Cytosol*
	3.1.3.5	*5'-nucleotidase*	NT5E	miR-193a-3p; miR-193b-3p; miR-30a-5p; miR-30b-5p; miR-30c-5p; miR-30d-5p; miR-30e-5p	*Plasma membrane;Cytosol*
			NT5C	ND	*Cytosol*
			NT5M	miR-124-3p.2; miR-506-3p	ND
			NT5C3A	miR-122-5p	ND
			NT5C2	miR-1-3p; miR-206	ND
	3.1.4.12	*sphingomyelin phosphodiesterase*	SMPD2	ND	*Vesicles;Plasma membrane;Cell Junctions*
			SMPD3	miR-124-3p.1; miR-128-3p; miR-133a-3p.1; miR-144-3p; miR-183-5p.1; miR-183-5p.2; miR-186-5p; miR-216a-5p; miR-218-5p; miR-219a-5p; miR-27a-3p; miR-27b-3p; miR-29a-3p; miR-29b-3p; miR-29c-3p; miR-504-5p.1; miR-599; miR-9-5p	*Endoplasmic reticulum*
			SMPD1	miR-15a-5p; miR-15b-5p; miR-16-5p; miR-195-5p; miR-424-5p; miR-497-5p	ND
			SMPD4	miR-133a-3p.2; miR-133b; miR-873-5p.2; miR-9-5p	*Nuclear membrane;Cytosol*
Liase	4.2.99.18	*DNA-(apurinic or apyrimidinic site) lyase*	NTHL1	ND	ND
			OGG1	ND	*Nucleoplasm*
			NEIL2	ND	*Nucleus;Nucleoli*
			NEIL1	ND	*Nucleoplasm*
			APEX1	miR-128-3p; miR-218-5p; miR-27a-3p; miR-27b-3p; miR-296-3p; miR-299-3p; miR-338-3p; miR-377-3p; miR-410-3p; miR-493-3p	*Nucleus*
			APEX2	miR-124-3p.1; miR-124-3p.2; miR-506-3p; miR-653-5p	*Nucleus;Nucleoli fibrillar center;Vesicles*
			APLF	miR-135a-5p; miR-135b-5p; miR-543	*Nucleoplasm*

De uma forma global, observando os perfis de transcrição dos genes em cada atividade enzimática, a distinção entre os perfis fica evidente, tanto entre formas análogas quanto entre parálogos. Levando em consideração apenas os tecidos nos quais a sobrexpressão dos genes foi constatada, observamos um padrão de alternância na expressão das formas análogas (e parálogas) como pode ser observado no Figura 4.4. Além disso, as predições computacionais de interação de miRNAs com os genes codificadores destas formas análogas revelaram grupos distintos de miRNAs possivelmente envolvidos na modulação da expressão de seus genes-alvos para 18 dos 31 genes do nosso conjunto de dados (58%), para os quais havia informação sobre miRNAs humanos e seus respectivos alvos no TargetScan (Tabela 4.2 e Figura 4.4); ainda, as informações de localização subcelular, experimentalmente obtidas, compiladas da base de dados *The Human Protein Atlas*, também mostram diferenças no local de atuação dos produtos de 21 genes do nosso conjunto de dados (67,7%), para os quais conseguimos informações neste repositório (Tabela 4.2).

Embora o resultado do agrupamento hierárquico tenha revelado, em alguns casos, similaridades entre os perfis de expressão de formas análogas intragenômicas, refletidas na proximidade entre estes genes no dendograma que representa o agrupamento, as distâncias medidas entre tais perfis são bastante significativas, conforme pode ser observado na Figura 4.4. Em média, as distâncias calculadas a partir da fórmula de correlação de Pearson modificada, entre os perfis de transcrição de formas análogas, para cada atividade enzimática foram: 0,54 (EC 1.15.1.1, superoxide dismutase); 0,48 (EC 2.7.1.67, 1-phosphatidylinositol 4-kinase); 0,69 (EC 3.1.2.2, palmitoyl-CoA hydrolase); 0,43 (3.1.3.2, acid phosphatase); 0,62 (EC 3.1.3.5, 5'-nucleotidase); 0,50 (EC 3.1.4.12, sphingomyelin phosphodiesterase); 0,63 (EC 4.2.99.18, DNA-(apurinic or apyrimidinic site) lyase). Por outro lado, as distâncias

médias entre os perfis transcricionais dos genes codificadores de formas parálogas neste mesmo grupo de atividades enzimáticas foram, respectivamente: 0,63; 0,65; 0,03; 0,57; 056; 0,49; 0,42.

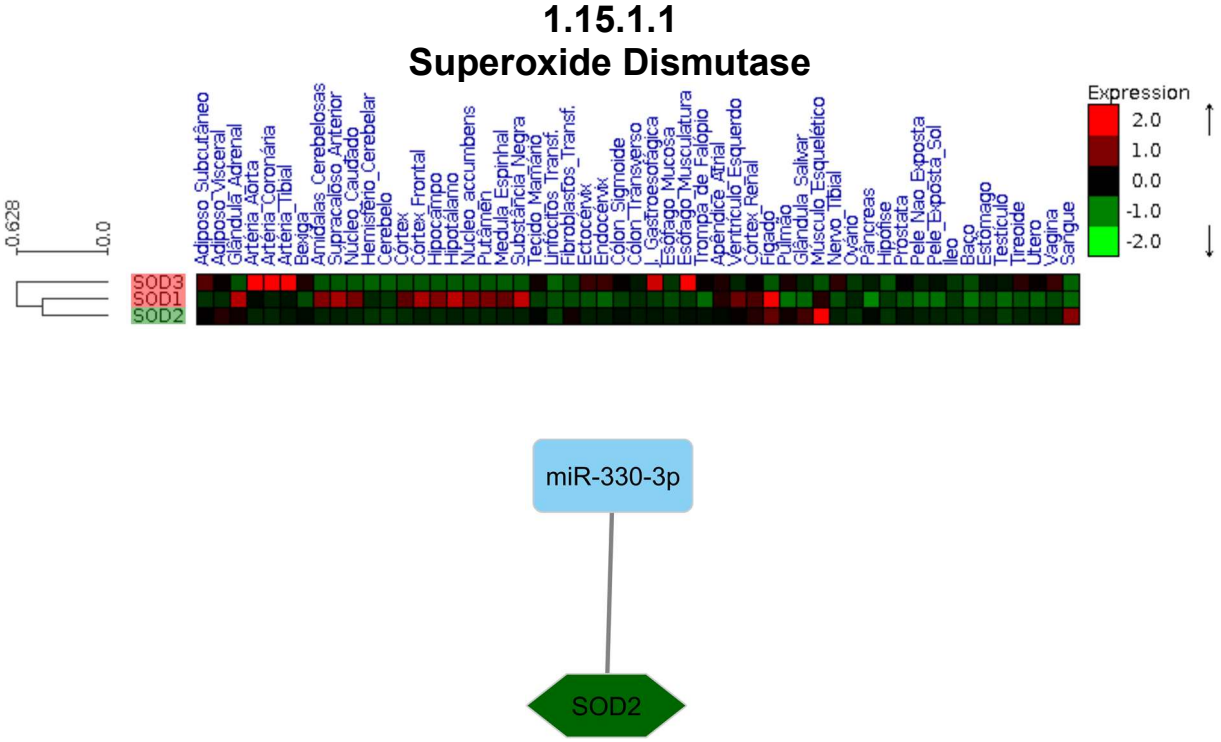


Figura 4.4 (Superior) Dendrogramas e matrizes (heatmaps) representando o resultado obtido com o agrupamento hierárquico dos perfis de expressão dos genes codificadores de formas análogas intragenômicas humanas. Com base nos valores de correlação, é produzido um heatmap com valores de -2 (verde) até 2 (vermelho), caracterizando para cada um dos 53 tecidos os genes com sub-expressão e sobre-expressão, respectivamente. (Inferior) Redes de interação dos miRNAs humanos e seus respectivos genes-alvos. As distintas enzimas isofuncionais não-homólogas de cada atividade enzimática estão representadas com as mesmas cores usadas na Tabela 4.2; miRNAs estão representados em azul e miRNAs com pelo menos dois alvos estão coloridos de cinza.

2.7.1.67
1-phosphatidylinositol 4-kinase

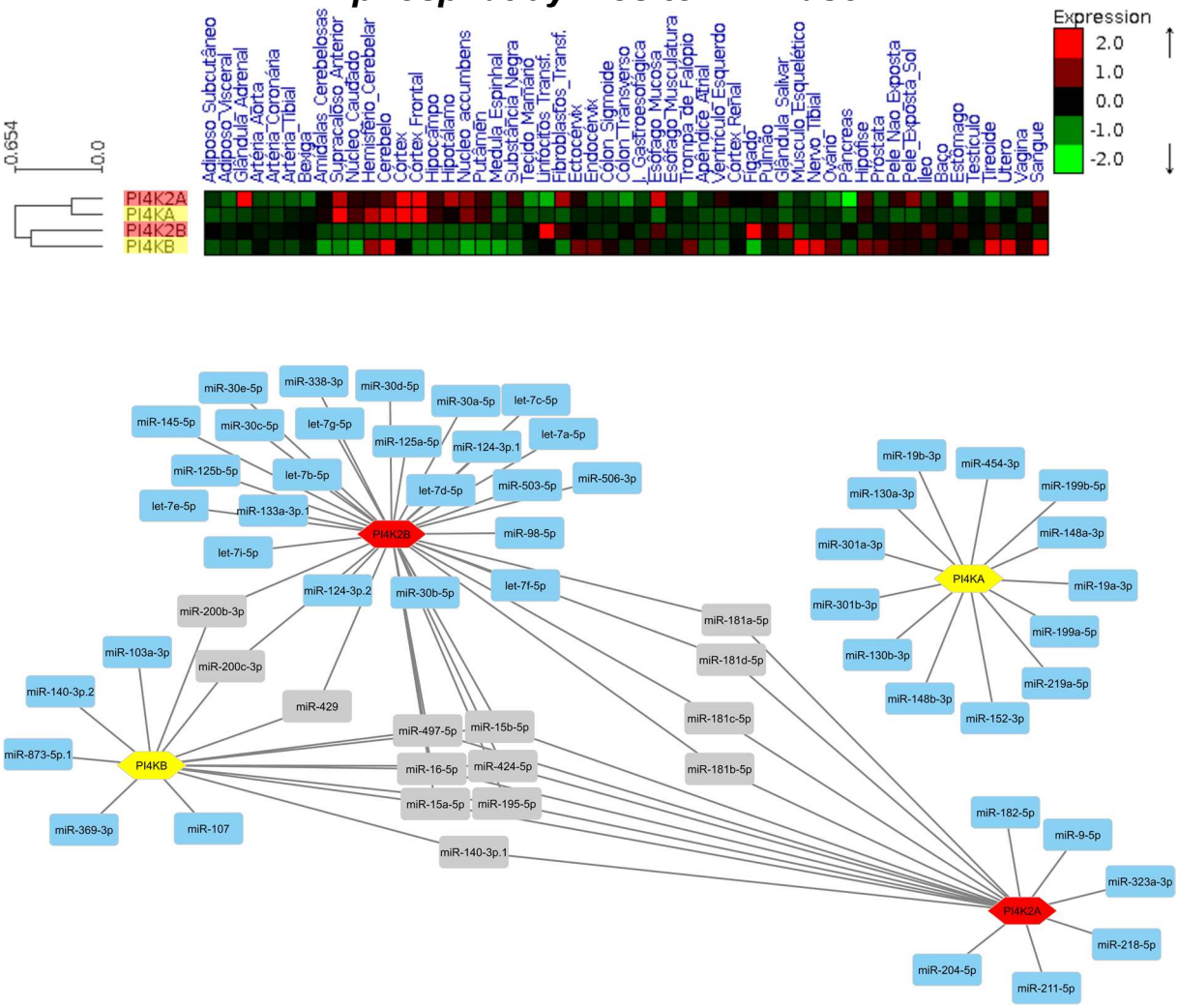


Figura 4.4

3.1.2.2
Palmitoyl-CoA Hydrolase

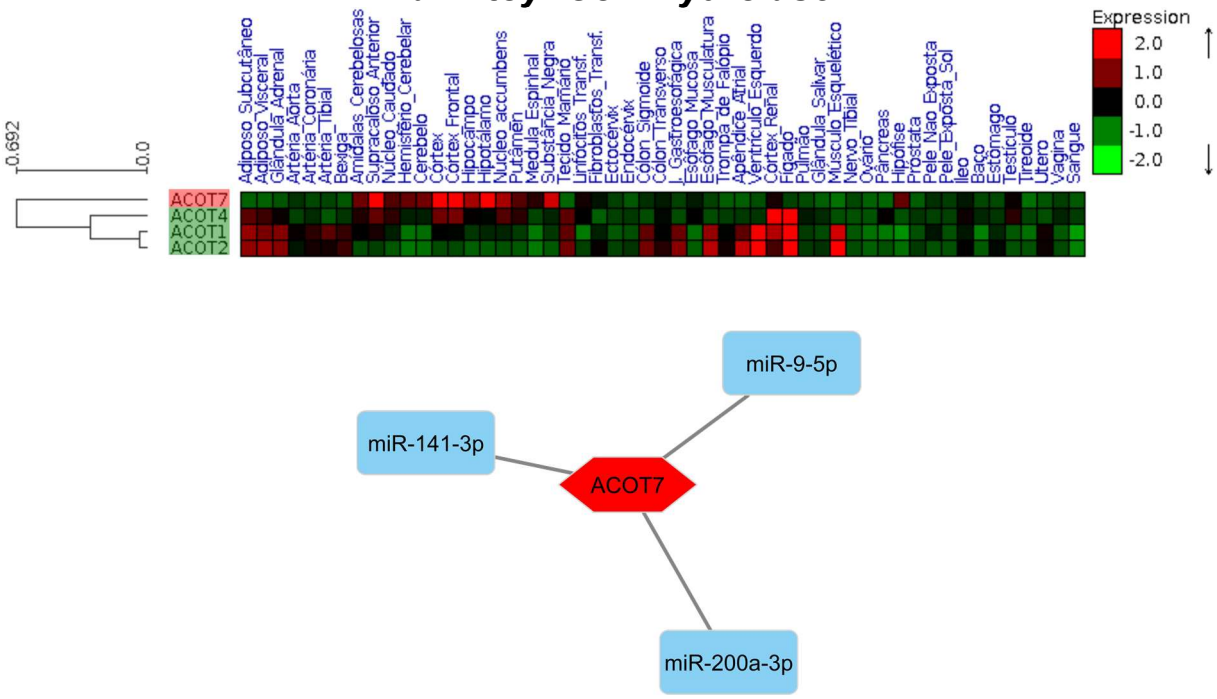


Figura 4.4

3.1.3.2
Acid Phosphatase

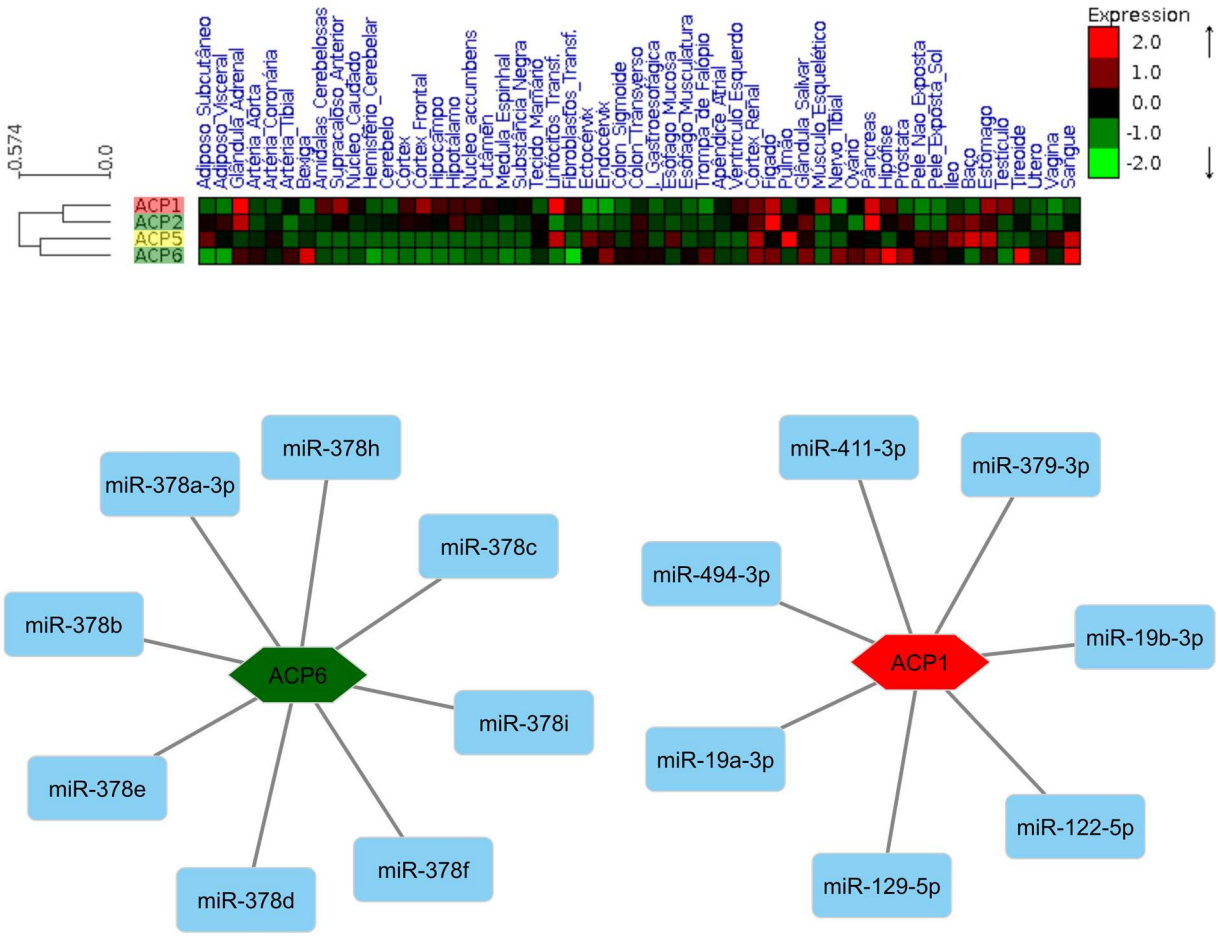


Figura 4.4

3.1.3.5
5'-nucleotidase

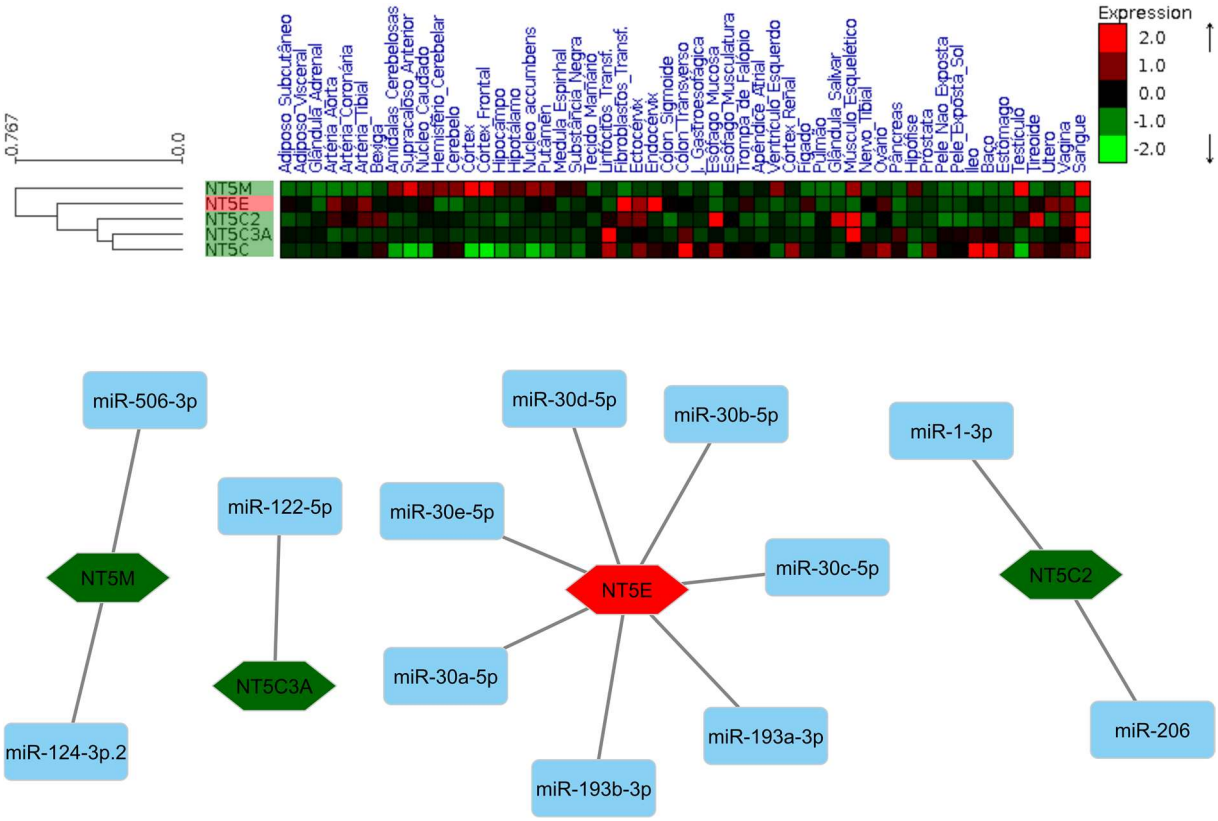


Figura 4.4

3.1.4.12
Sphingomyelin Phosphodiesterase

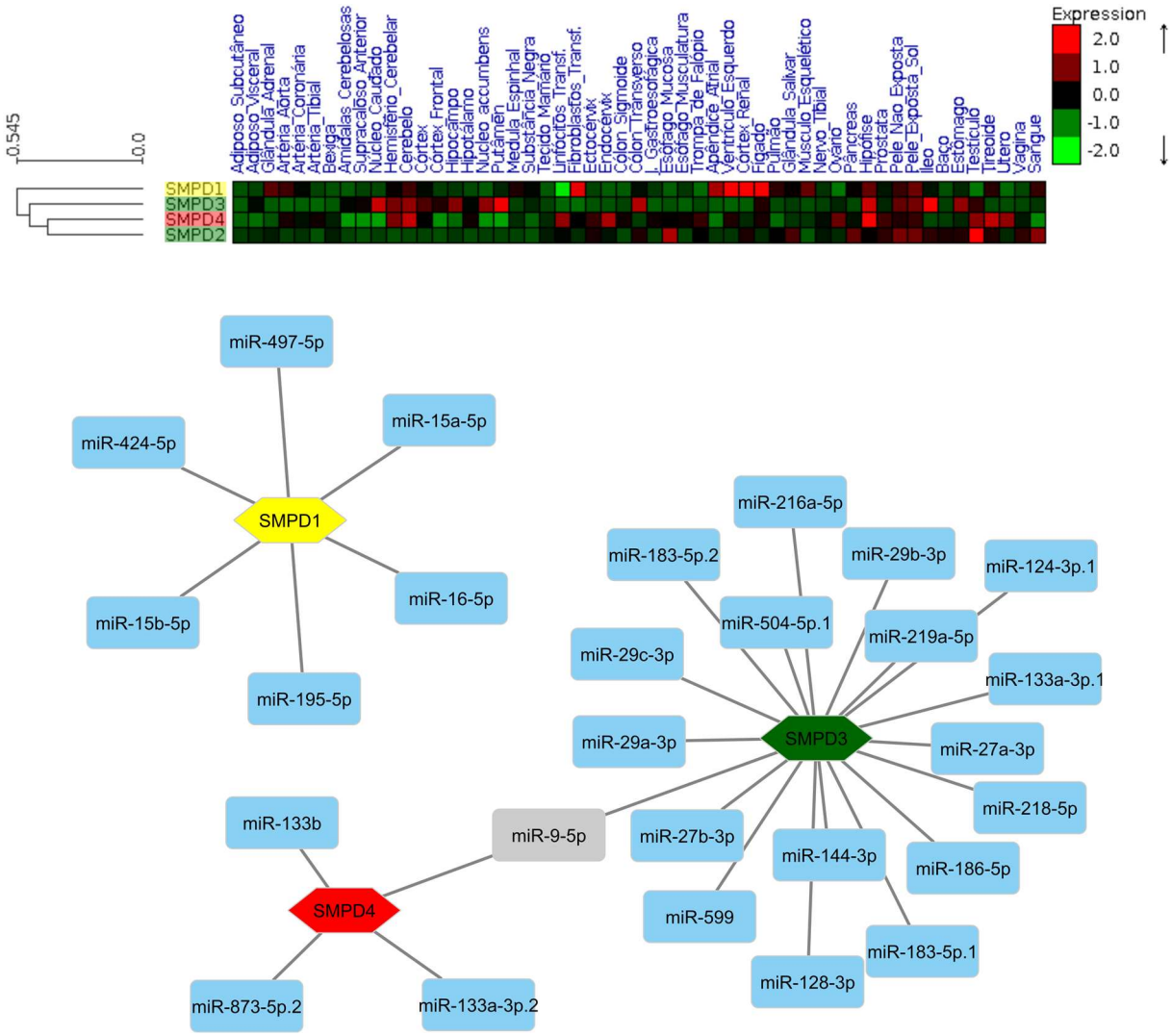


Figura 4.4

4.2.99.18
DNA-(apurinic or apyrimidinic site) lyase

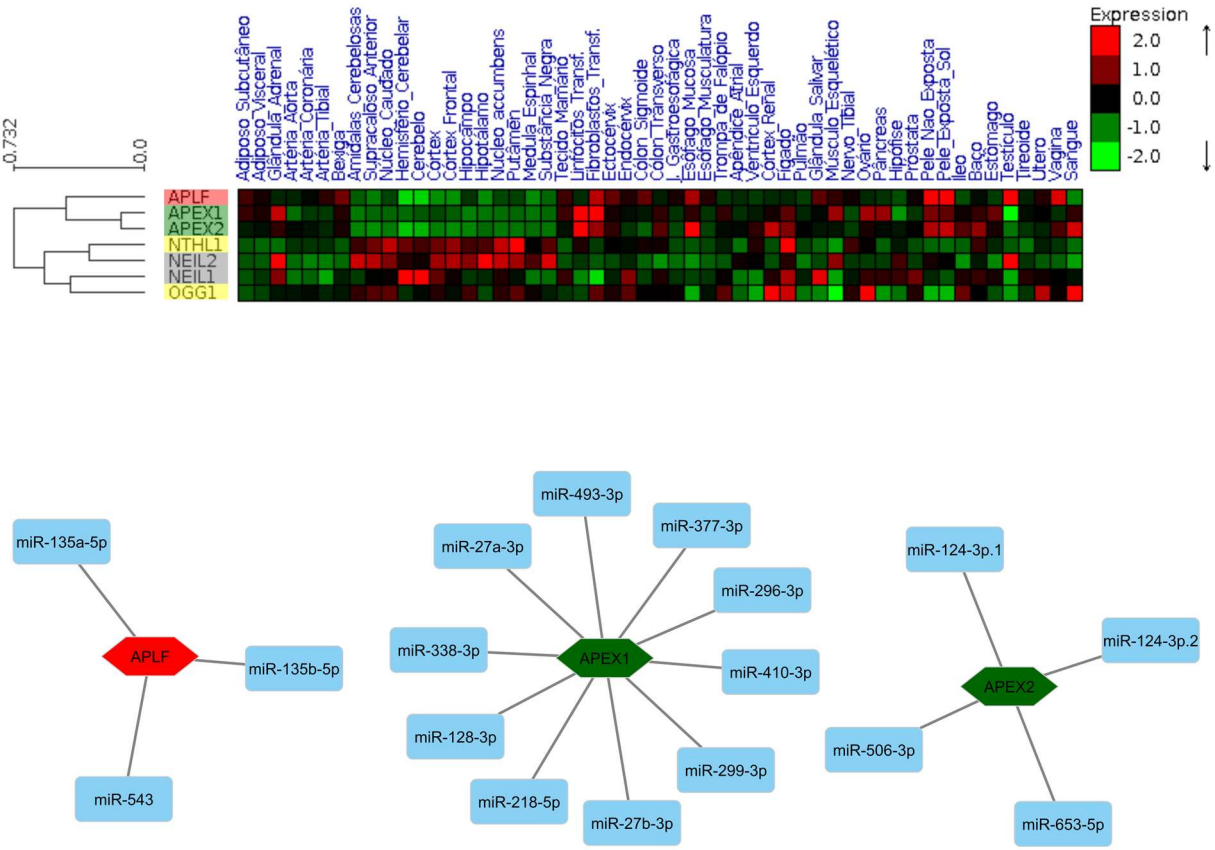


Figura 4.4

1.15.1.1

	SOD2	SOD1	SOD3
SOD2	0		
SOD1	0.44988555	0	
SOD3	0.62821454	0.62821454	0

2.7.1.67

	PI4KA	PI4KB	PI4K2A	PI4K2B
PI4KA	0			
PI4KB	0.6545496	0		
PI4K2A	0.2247912	0.6545496	0	
PI4K2B	0.5286664	0.5286664	0.6545496	0

3.1.2.2

	ACOT2	ACOT1	ACOT4	ACOT7
ACOT2	0			
ACOT1	0.03697923	0		
ACOT4	0.2958235	0.2958235	0	
ACOT7	0.69274056	0.69274056	0.69274056	0

3.1.3.2

	ACP5	ACP2	ACP6	ACP1
ACP5	0			
ACP2	0.57456875	0		
ACP6	0.4381461	0.57456875	0	
ACP1	0.57456875	0.29645053	0.57456875	0

3.1.3.5

	NT5E	NT5C	NT5M	NT5C3A	NT5C2
NT5E	0				
NT5C	0.57340527	0			
NT5M	0.7676202	0.7676202	0		
NT5C3A	0.57340527	0.31497723	0.7676202	0	
NT5C2	0.57340527	0.38626796	0.7676202	0.38626796	0

3.1.4.12

	SMPD2	SMPD3	SMPD1	SMPD4
SMPD2	0			
SMPD3	0.4903009	0		
SMPD1	0.5452814	0.5452814	0	
SMPD4	0.40620843	0.4903009	0.5452814	0

4.2.99.18

	NTHL1	OGG1	NEIL2	NEIL1	APEX1	APEX2	APLF
NTHL1	0						
OGG1	0.56009966	0					
NEIL2	0.30259183	0.56009966	0				
NEIL1	0.56009966	0.40976486	0.56009966	0			
APEX1	0.7321852	0.7321852	0.7321852	0.7321852	0		
APEX2	0.7321852	0.7321852	0.7321852	0.7321852	0.12625223	0	
APLF	0.7321852	0.7321852	0.7321852	0.7321852	0.43297595	0.43297595	0

Figura 4.5 Matrizes representando as distâncias estimadas durante o agrupamento hierárquico aglomerativo ao qual o conjunto de dados descrito na Tabela 4.2 fui submetido. Estão representadas as 7 atividades enzimáticas estudadas, com os genes codificadores de enzimas isofuncionais não-homólogas intragenômicas representados com cores diferentes (seguindo o mesmo padrão usado na Tabela 4.2).. A repetição exata de determinados números corresponde às distâncias medidas entre um agrupamento de genes e outros genes ou grupos de genes, de acordo com o dendrograma correspondente (Figura 4.4).

Comparando as distâncias médias calculadas entre os perfis de transcrição de análogos e parálogos, em cada atividade enzimática, a única diferença significativa observada ocorre na atividade EC 3.1.2.2 (palmitoyl-CoA hydrolase), na qual as distâncias entre as formas análogas foi de 0,69 (ACOT7 *versus* ACOT2, ACOT1 e ACOT4) ao passo que entre as formas parálogas foi de 0,018 (ACOT2, ACOT1 e ACOT4 entre si). Levando em consideração o grau de similaridade medido (alinhamento global rigoroso) entre as enzimas codificadas pelas formas parálogas ACOT2 e ACOT1 (86,5%), ACOT2 e ACOT4 (70,6%) e ACOT1 e ACOT4 (81%), a distribuição filogenética de tais genes homólogos (todos pertencentes ao mesmo grupo de ortólogos K01068), conservados entre os grupos de eucariontes fungos, plantas e animais, bem como a localização subcelular (as enzimas ACOT1 e ACOT2 são expressas nas mitocôndrias) é de se esperar que seus produtos sejam igualmente conservados do ponto de vista funcional. Todos os demais grupos de parálogos, em todas as atividades analisadas em nossa pesquisa apresentaram similaridade abaixo de 35% quando tiveram suas cadeias de aminoácidos alinhadas (Tabela 4.1), com excessão de dois casos apenas: PI4K2A e PI4K2B (EC 3.1.2.2), enzimas com localização subcelular na membrana plasmática e no citoplasma, respectivamente (Tabela 4.2), compartilhando 69,5% de similaridade entre si; e NT5C e NT5M (EC 3.1.3.5), possivelmente transferidos horizontalmente para a linhagem humana, conforme discutido anteriormente, apresentando 64,7% de similaridade entre si.

Finalmente, comparando os genes parálogos APEX1 e APEX2 (EC 4.2.99.18) (Figura 4.5), constata-se uma similaridade de 88% (distância de 0.12) entre seus perfis de expressão, apesar da baixa similaridade em nível de sequência (22%) (Tabela 4.1) e de serem expressos em distintos locais da célula: núcleo (APEX1); núcleo, vesícula e nucléolo (APEX2). Vale destacar que, segundo os mapas de processos celulares de referência do KEGG, esses genes são anotados como componentes do complexo de

reparo por excisão de base (BER – *base excision repair*) e, ainda, a conservação no perfil de expressão de parálogos, como observado entre APEX1 e APEX2, pode refletir genes que cooperam dentro de uma mesma via metabólica ou na arquitetura de complexos proteicos (108).

5 CONCLUSÕES

Neste trabalho, encontramos evidências de enzimas isofuncionais não-homólogas catalisando em 15 atividades enzimáticas (compreendendo 70 sequências enzimáticas no total) do metabolismo humano. Notavelmente, apesar do uso de critérios muito restritivos (excluindo enzimas multiméricas, atividades enzimáticas com classificação incompleta de EC, bem como grupos compostos exclusivamente por uma única sequência humana) e nosso foco em atividades enzimáticas humanas nas quais a participação de enzimas não relacionadas possa ser reconhecida, descobrimos enzimas análogas intragenômicas em 3 atividades enzimáticas (20% de nosso conjunto de dados *bona fide*) sem nenhuma evidência de analogia relatada até agora: *lactose synthase* (EC 2.4.1.22), *NAD+—protein-arginine ADP-ribosyltransferase* (EC 2.4.2.31) e *1-phosphatidylinositol 4-kinase* (EC 2.7.1.67). Essas atividades enzimáticas participam de 9 vias bioquímicas ou processos biológicos distintos, alguns dos quais desempenham papéis essenciais no câncer, metabolismo da galactose, biossíntese de glicosaminoglicanos, biossíntese de glicoesfingolipídios, metabolismo de inositol fosfato, biossíntese de manose O-glicanos, biossíntese de n-glicanos, biossíntese de outros tipos de O-glicanos, e do sistema de sinalização do fosfatidilinositol.

Os resultados das análises comparativas dos perfis de expressão, redes de interação com miRNAs e localização subcelular parecem refutar a hipótese de que a coexistência de enzimas isofuncionais não-homólogas na espécie humana se caracteriza como redundância funcional, uma vez que, provavelmente, essas proteínas estão envolvidas em contextos biológicos distintos. O perfil de expressão dos genes codificadores de enzimas análogas revelou também alternância na expressão no perfil nos 53 tecidos estudados.

Além disso, ficou evidente que as informações sobre todas as enzimas necessitam de uma indicação clara de suas origens evolutivas de uma forma muita mais refinada do que as utilizadas nas bases de dados atuais. O sistema de ortologia do KEGG (KO; *KEGG Orthology*), por exemplo, é empregado na construção de vias metabólicas (além de outros módulos da plataforma). Esse tipo de generalização impacta no estudo das vias metabólicas, não só pelo ponto de vista evolutivo, mas também prático. A atribuição de informações evolutivas para cada atividade enzimática (análogos funcionais) poderá auxiliar na busca de novos alvos terapêuticos e até mesmo para outras finalidades.

6 REFERÊNCIAS BIBLIOGRÁFICAS

1. Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, Darnell J. Molecular Cell Biology. 4th edition. New York: W. H. Freeman. 2000.
2. Kolodny R, Pereyaslavets L, Samson AO, Levitt M. On the Universe of Protein Folds. Annu Rev Biophys. 2013;
3. Berg J, Tymoczko J, Stryer L. Biochemistry. New York. 2002.
4. Drozdetskiy A, Cole C, Procter J, Barton GJ. JPred4: A protein secondary structure prediction server. Nucleic Acids Res. 2015;
5. Hasegawa H, Holm L. Advances and pitfalls of protein structural alignment. Current Opinion in Structural Biology. 2009.
6. Godzik A. The structural alignment between two proteins: Is there a unique answer? Protein Sci. 1996;
7. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res [Internet]. 2005 Apr 11;33(7):2302–9. Available from: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gki524
8. Levitt M, Gerstein M. A unified statistical framework for sequence comparison and structure comparison. Proc Natl Acad Sci USA. 1998;
9. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol [Internet]. 1995 Apr 7;247(4):536–40. Available from: http://www.ncbi.nlm.nih.gov/pubmed/7723011
10. Lo Conte L, Ailey B, Hubbard TJ, Brenner SE, Murzin AG, Chothia C. SCOP: a structural classification of proteins database. Nucleic Acids Res [Internet]. 2000 Jan 1;28(1):257–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10592240
11. George RA, Spriggs R V., Thornton JM, Al-Lazikani B, Swindells MB. SCOPEC: a database of protein catalytic domains. Bioinformatics [Internet]. 2004 Aug 4;20(Suppl 1):i130–6. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bth948
12. Meister A, Anderson ME. Glutathione. Annu Rev Biochem [Internet]. 1983 Jun;52(1):711–60. Available from: http://www.annualreviews.org/doi/10.1146/annurev.bi.52.070183.003431
13. Wijayanti N, Katz N, Immenschuh S. Biology of heme in health and disease. Curr

Med Chem [Internet]. 2004 Apr;11(8):981–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15078160

14. Linster CL, Van Schaftingen E. Vitamin C. Biosynthesis, recycling and degradation in mammals. FEBS J [Internet]. 2007 Jan;274(1):1–22. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17222174
15. McDonald AG, Tipton KF. Fifty-five years of enzyme classification: advances and difficulties. FEBS J [Internet]. 2014 Jan;281(2):583–92. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24103004
16. Forslund K, Sonnhammer ELL. Evolution of protein domain architectures. Methods Mol Biol [Internet]. 2012;856:187–216. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22399460
17. Stein L. Genome annotation: From sequence to biology. Nature Reviews Genetics. 2001.
18. Gould SJ. Evolution and the Triumph of Homology, or why History Matters. Am Sci. 1986;
19. Hall BK. Homoplasy and homology: Dichotomy or continuum? J Hum Evol. 2007;
20. Lockwood CA, Fleagle JG. Homoplasy in primate and human evolution. Journal of Human Evolution. 2007;
21. Rendall D, Di Fiore A. Homoplasy, homology, and the perceived special status of behavior in evolution. J Hum Evol. 2007;
22. Fitch WM. Homology a personal view on some of the problems. Trends Genet. 2000;
23. Glaeser G, Paulus HF, Nachtigall W. The Evolution of Flight. The Evolution of Flight. 2017.
24. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, et al. Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci. 2007;
25. Thiele I, Swainston N, Fleming RMT, Hoppe A, Sahoo S, Aurich MK, et al. A community-driven global reconstruction of human metabolism. Nat Biotechnol [Internet]. 2013;31(5):419–25. Available from: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=23455439&retmode=ref&cmd=prlinks
26. Cordwell SJ. Microbial genomes and “missing” enzymes: redefining biochemical pathways. Arch Microbiol [Internet]. 1999 Oct 14;172(5):269–79. Available from: http://link.springer.com/10.1007/s002030050780

27. Galperin MY, Koonin E V. Functional genomics and enzyme evolution. Homologous and analogous enzymes encoded in microbial genomes. Genetica [Internet]. 1999;106(1–2):159–70. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10710722
28. Huynen M a, Dandekar T, Bork P. Variation and evolution of the citric-acid cycle: a genomic perspective. Trends Microbiol [Internet]. 1999 Jul;7(7):281–91. Available from: isi:000081379900011
29. Peregrin-Alvarez JM, Tsoka S, Ouzounis CA. The phylogenetic extent of metabolic enzymes and pathways. Genome Res [Internet]. 2003 Mar 1;13(3):422–7. Available from: http://www.genome.org/cgi/doi/10.1101/gr.246903
30. Hanson AD, Pribat A, Waller JC, Crécy-Lagard Valérie de. 'Unknown' proteins and 'orphan' enzymes: the missing half of the engineering parts list – and how to find it. Biochem J. 2010;
31. Almonacid DE, Yera ER, Mitchell JBO, Babbitt PC. Quantitative Comparison of Catalytic Mechanisms and Overall Reactions in Convergently Evolved Enzymes: Implications for Classification of Enzyme Function. Orengo CA, editor. PLoS Comput Biol [Internet]. 2010 Mar 12;6(3):e1000700. Available from: http://dx.plos.org/10.1371/journal.pcbi.1000700
32. Galperin MY, Koonin E V. Divergence and convergence in enzyme evolution. J Biol Chem. 2012;287(1):21–8.
33. Galperin MY, Walker DR, Koonin E V. Analogous enzymes: independent inventions in enzyme evolution. Genome Res [Internet]. 1998 Aug;8(8):779–90. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9724324
34. Omelchenko M V, Galperin MY, Wolf YI, Koonin E V. Non-homologous isofunctional enzymes: a systematic analysis of alternative solutions in enzyme evolution. Biol Direct [Internet]. 2010 Apr 30;5(1):31. Available from: http://www.biology-direct.com/content/5/1/31
35. Doolittle RF. Convergent evolution: the need to be explicit. Trends Biochem Sci [Internet]. 1994 Jan;19(1):15–8. Available from: http://linkinghub.elsevier.com/retrieve/pii/0968000494901678
36. Gherardini PF, Wass MN, Helmer-Citterich M, Sternberg MJE. Convergent evolution of enzyme active sites is not a rare phenomenon. J Mol Biol [Internet]. 2007 Sep 21;372(3):817–45. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17681532

37. Otto TD, Guimarães ACR, Degrave WM, de Miranda AB. AnEnPi: identification and annotation of analogous enzymes. BMC Bioinformatics [Internet]. 2008 Dec 17;9:544. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19091081
38. Gomes MR, Guimarães ACR, de Miranda AB. Specific and nonhomologous isofunctional enzymes of the genetic information processing pathways as potential therapeutical targets for tritryps. Enzyme Res. 2011;2011:543912.
39. Alves-Ferreira M, Guimarães ACR, Capriles PV da SZ, Dardenne LE, Degrave WM, Guimaraes ACR, et al. A new approach for potential drug target discovery through in silico metabolic pathway analysis using *Trypanosoma cruzi* genome information. Mem Inst Oswaldo Cruz [Internet]. 2009;104(8):1100–10. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20140370
40. Guimarães AC, Otto TD, Alves-Ferreira M, Miranda AB, Degrave WM. In silico reconstruction of the amino acid metabolic pathways of Trypanosoma cruzi. Genet Mol Res. 2008;7(3):872–82.
41. Capriles PVSZ, Guimarães ACR, Otto TD, Miranda AB, Dardenne LE, Degrave WM. Structural modelling and comparative analysis of homologous, analogous and specific proteins from *Trypanosoma cruzi* versus *Homo sapiens*: putative drug targets for chagas' disease treatment. BMC Genomics. 2010;11:610.
42. Pettersson E, Lundeberg J, Ahmadian A. Generations of sequencing technologies. Genomics. 2009.
43. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature [Internet]. 2001;409(6822):860–921. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11237011%5Cnhttp://www.nature.com/nature/journal/v409/n6822/pdf/409860a0.pdf
44. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science [Internet]. 2001;291(5507):1304–51. Available from: http://www.sciencemag.org/content/291/5507/1304.abstract
45. Reddy TBK, Thomas AD, Stamatis D, Bertsch J, Isbandi M, Jansson J, et al. The Genomes OnLine Database (GOLD) v.5: A metadata management system based on a four level (meta)genome project classification. Nucleic Acids Res. 2015;43(D1):D1099–106.
46. Ansorge WJ. Next-generation DNA sequencing techniques. New Biotechnology. 2009.
47. Kircher M, Kelso J. High-throughput DNA sequencing--concepts and limitations.

Bioessays. 2010;

48. Elgar G, Vavouri T. Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. Trends in Genetics. 2008.
49. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science [Internet]. 2004 Oct 22;306(5696):636–40. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15499007
50. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: The reference human genome annotation for the ENCODE project. Genome Res. 2012;
51. Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. Nature Reviews Genetics. 2009.
52. Yang IS, Kim S. Analysis of Whole Transcriptome Sequencing Data: Workflow and Software. Genomics Inform [Internet]. 2015 Dec;13(4):119–25. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26865842
53. Han Y, Gao S, Muegge K, Zhang W, Zhou B. Advanced applications of RNA sequencing and challenges. Bioinform Biol Insights. 2015;
54. Hrdlickova R, Toloue M, Tian B. RNA-Seq methods for transcriptome analysis. Wiley Interdisciplinary Reviews: RNA. 2017.
55. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res [Internet]. 2008 Sep;18(9):1509–17. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18550803
56. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. Nat Biotechnol [Internet]. 2014 Sep;32(9):896–902. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25150836
57. Tjaden B. De novo assembly of bacterial transcriptomes from RNA-seq data. Genome Biol [Internet]. 2015 Jan 13;16:1. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25583448
58. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17(1).
59. Aravin AA, Lagos-Quintana M, Yalcin A, Zavolan M, Marks D, Snyder B, et al. The small RNA profile during Drosophila melanogaster development. Dev Cell [Internet]. 2003 Aug;5(2):337–50. Available from:

http://www.ncbi.nlm.nih.gov/pubmed/12919683

60. Gulyaeva LF, Kushlinskiy NE. Regulatory mechanisms of microRNA expression. Journal of Translational Medicine. 2016.
61. Maltby S, Plank M, Tay HL, Collison A, Foster PS. Targeting MicroRNA function in respiratory diseases: Mini-review. Frontiers in Physiology. 2016.
62. Wang WX, Danaher RJ, Miller CS, Berger JR, Nubia VG, Wilfred BS, et al. Expression of miR-15/107 family microRNAs in human tissues and cultured rat brain cells. Genomics, Proteomics Bioinforma. 2014;
63. Ambros V, Bartel B, Bartel DP, Burge CB, Carrington JC, Chen X, et al. A uniform system for microRNA annotation. RNA. 2003;
64. Carthew RW, Sontheimer EJ. Origins and Mechanisms of miRNAs and siRNAs. Cell. 2009.
65. Martin W, Schnarrenberger C. The evolution of the Calvin cycle from prokaryotic to eukaryotic chromosomes: a case study of functional redundancy in ancient pathways through endosymbiosis. Curr Genet [Internet]. 1997 Jul;32(1):1–18. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9309164
66. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res [Internet]. 2000 Jan 1;28(1):27–30. Available from: http://nar.oxfordjournals.org/cgi/content/abstract/28/1/27
67. Fernandes A da F. ANENDB : PREDIÇÃO COMPUTACIONAL E BANCO DE DADOS PARA ENZIMAS ANÁLOGAS. Fundação Oswaldo Cruz; 2016.
68. Altschul S. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res [Internet]. 1997 Sep 1;25(17):3389–402. Available from: http://dx.doi.org/10.1093/nar/25.17.3389
69. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res [Internet]. 2014 Jan;42(D1):D222–30. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkt1223
70. Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, Madera M, et al. SUPERFAMILY--sophisticated comparative genomics, data mining, visualization and phylogeny. Nucleic Acids Res [Internet]. 2009 Jan;37(Database issue):D380-6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19036790%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2686452
71. Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open

Software Suite. Trends Genet. 2000;16(1):276–7.

72. Webb B, Sali A. Comparative Protein Structure Modeling Using MODELLER. Curr Protoc Bioinforma [Internet]. 2014 Sep 8;47:5.6.1-32. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25199792

73. Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, et al. MolProbity : all-atom structure validation for macromolecular crystallography. Acta Crystallogr Sect D Biol Crystallogr [Internet]. 2010 Jan 1;66(1):12–21. Available from: http://scripts.iucr.org/cgi-bin/paper?S0907444909042073

74. Chen VB, Davis IW, Richardson DC. KING (Kinemage, Next Generation): A versatile interactive molecular and scientific visualization program. Protein Sci [Internet]. 2009 Nov;18(11):2403–9. Available from: http://doi.wiley.com/10.1002/pro.250

75. Xu D, Zhang Y. Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. Biophys J [Internet]. 2011 Nov 16;101(10):2525–34. Available from: http://dx.doi.org/10.1016/j.bpj.2011.10.024

76. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins [Internet]. 2004 Dec 1;57(4):702–10. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15476259

77. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2015. Nucleic Acids Res [Internet]. 2015 Jan 28;43(D1):D662–9. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gku1010

78. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res [Internet]. 2009 Sep;19(9):1639–45. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19541911

79. Consortium TGte. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45(6):580–5.

80. Melé M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, et al. The human transcriptome across tissues and individuals. Science (80-). 2015;

81. Trapnell C, Pachter L, Salzberg SL. TopHat: Discovering splice junctions with RNA-Seq. Bioinformatics. 2009;

82. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;

83. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol [Internet]. 2004;5(10):R80. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15461798
84. Sharan R, Maron-Katz A, Shamir R. CLICK and EXPANDER: a system for clustering and visualizing gene expression data. Bioinformatics [Internet]. 2003 Sep 22;19(14):1787–99. Available from: http://www.ncbi.nlm.nih.gov/pubmed/14512350
85. Ulitsky I, Maron-Katz A, Shavit S, Sagir D, Linhart C, Elkon R, et al. Expander: from expression microarrays to networks and functions. Nat Protoc [Internet]. 2010 Feb;5(2):303–22. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20134430
86. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell. 2005;
87. Grimson A, Farh KKH, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP. MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing. Mol Cell. 2007;
88. Friedman RC, Farh KKH, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. Genome Res. 2009;
89. Garcia DM, Baek D, Shin C, Bell GW, Grimson A, Bartel DP. Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. Nat Struct Mol Biol [Internet]. 2011 Sep 11;18(10):1139–46. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21909094
90. Thul PJ, Åkesson L, Wiking M, Mahdessian D, Geladaki A, Ait Blal H, et al. A subcellular map of the human proteome. Science (80-). 2017;
91. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2.8: New features for data integration and network visualization. Bioinformatics. 2011;
92. Sillitoe I, Lewis TE, Cuff A, Das S, Ashford P, Dawson NL, et al. CATH: Comprehensive structural and functional annotations for genome sequences. Nucleic Acids Res. 2015;
93. Crisp A, Boschetti C, Perry M, Tunnacliffe A, Micklem G. Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. Genome Biol [Internet]. 2015;16(1):50. Available from: http://genomebiology.com/2015/16/1/50

94. Zimmermann H. 5’-Nucleotidase: molecular structure and functional aspects. Biochem J. 1992;285 (Pt 2:345–65.
95. Penning TM, Chen M, Jin Y. Promiscuity and diversity in 3-ketosteroid reductases. Journal of Steroid Biochemistry and Molecular Biology. 2015.
96. Urade Y, Eguchi N. Lipocalin-type and hematopoietic prostaglandin D synthases as a novel example of functional convergence. Prostaglandins Other Lipid Mediat. 2002;
97. Lim SM, Chen D, Teo H, Roos A, Jansson AE, Nyman T, et al. Structural and dynamic insights into substrate binding and catalysis of human lipocalin prostaglandin D synthase. J Lipid Res [Internet]. 2013 Jun 1 [cited 2016 Apr 5];54(6):1630–43. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3646464&tool=pmcentrez&rendertype=abstract
98. Trimarco A, Forese MG, Alfieri V, Lucente A, Brambilla P, Dina G, et al. Prostaglandin D2 synthase/GPR44: a signaling axis in PNS myelination. Nat Neurosci [Internet]. 2014 Dec;17(12):1682–92. Available from: http://www.nature.com/neuro/journal/v17/n12/full/nn.3857.html%5Cnhttp://www.nature.com/neuro/journal/v17/n12/pdf/nn.3857.pdf
99. Bridges PJ, Jeoung M, Shim S, Park JY, Lee JE, Sapsford LA, et al. Hematopoetic Prostaglandin D Synthase: An ESR1-Dependent Oviductal Epithelial Cell Synthase. Endocrinology [Internet]. 2012 Apr;153(4):1925–35. Available from: http://press.endocrine.org/doi/10.1210/en.2011-1900
100. Marín-Méndez JJ, Patiño-García A, Segura V, Ortuño F, Gálvez MD, Soutullo CA. Differential expression of prostaglandin D2 synthase (PTGDS) in patients with attention deficit-hyperactivity disorder and bipolar disorder. J Affect Disord [Internet]. 2012 May;138(3):479–84. Available from: http://dx.doi.org/10.1016/j.jad.2012.01.040
101. Tanaka K, Ogawa K, Sugamura K, Nakamura M, Takano S, Nagata K. Cutting edge: differential production of prostaglandin D2 by human helper T cell subsets. J Immunol [Internet]. 2000 Mar 1;164(5):2277–80. Available from: http://www.jimmunol.org/cgi/content/abstract/164/5/2277
102. Landis GN, Tower J. Superoxide dismutase evolution and life span regulation. Mech Ageing Dev [Internet]. 2005 Mar;126(3):365–79. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0047637404001733
103. Garcia YM, Barwinska-Sendra A, Tarrant E, Skaar EP, Waldron KJ, Kehl-Fie TE.

A Superoxide Dismutase Capable of Functioning with Iron or Manganese Promotes the Resistance of *Staphylococcus aureus* to Calprotectin and Nutritional Immunity. Peschel A, editor. PLOS Pathog [Internet]. 2017 Jan 19;13(1):e1006125. Available from: http://dx.plos.org/10.1371/journal.ppat.1006125

104. Barylko B, Gerber SH, Binns DD, Grichine N, Khvotchev M, Südhof TC, et al. A novel family of phosphatidylinositol 4-kinases conserved from yeast to humans. J Biol Chem [Internet]. 2001 Mar 16;276(11):7705–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11244087
105. Heilmeyer LMG, Vereb G, Vereb G, Kakuk A, Szivák I. Mammalian phosphatidylinositol 4-kinases. IUBMB Life [Internet]. 2003 Feb;55(2):59–65. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12749687
106. Boura E, Nencka R. Phosphatidylinositol 4-kinases: Function, structure, and inhibition. Exp Cell Res [Internet]. 2015 Oct 1;337(2):136–45. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26183104
107. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. 1998;
108. Ibn-Salem J, Muro EM, Andrade-Navarro MA. Co-regulation of paralog genes in the three-dimensional chromatin architecture. Nucleic Acids Res. 2017;
109. Liu W, Xie Y, Ma J, Luo X, Nie P, Zuo Z, et al. Sequence analysis IBS : an illustrator for the presentation and visualization of biological sequences. 2015;31(June):3359–61.

ANEXOS

Anexo I: Características dos 70 genes codificadores de enzimas com evidência de analogia funcional das 15 atividades enzimáticas que apresentavam dois conjuntos distintos de superfamílias. Foram representados para cada enzima: as classes enzimáticas; os ECs; o nome que descreve a reação; as vias metabólicas; seus genes; agrupamento atribuído pelo AnEnPi; grupo de ortólogos do KEGG (KO); informações estruturais obtidas do PDB; anotações de superfamílias (SUPERFAMILY); atribuições de domínios funcionais e famílias (Pfam).

Classe	EC	Atividade	Via Metabólica	Gene	AnEnPi	KO	PDB	SUPERFAMILY	Pfam
Oxidorredutase	1.15.1.1	*superoxide dismutase*	*Huntington's disease* *Longevity regulating pathway* *Peroxisome* *Prion diseases* *Amyotrophic lateral sclerosis (ALS)* *FoxO signaling pathway*	SOD2	1	K04564	1LUV	SSF46609 SSF54719	PF00081 PF02777
Oxidorredutase	1.15.1.1	*superoxide dismutase*	*Huntington's disease* *Longevity regulating pathway* *Peroxisome* *Prion diseases* *Amyotrophic lateral sclerosis (ALS)* *FoxO signaling pathway*	SOD1	2	K04565	4XCR	SSF49329	PF00080
Oxidorredutase	1.15.1.1	*superoxide dismutase*	*Huntington's disease* *Longevity regulating pathway* *Peroxisome* *Prion diseases* *Amyotrophic lateral sclerosis (ALS)* *FoxO signaling pathway*	SOD3	2	K16627	2JLP	SSF49329	PF00080
Oxidorredutase	1.3.1.20	*trans-1,2-dihydrobenzene-1,2-diol dehydrogenase*	*Metabolism of xenobiotics by cytochrome P450* *Pentose and glucuronate interconversions* *Steroid hormone biosynthesis*	DHDH	1	K00078	2O48	SSF51735 SSF55347	PF01408

Classe	EC	Atividade	Via Metabólica	Gene	AnEnPi	KO	PDB	SUPERFAMILY	Pfam
Oxidorredutase	**1.3.1.20**	***trans-1,2-dihydrobenzene-1,2-diol dehydrogenase***	***Metabolism of xenobiotics by cytochrome P450 Pentose and glucuronate interconversions Steroid hormone biosynthesis***	**AKR1C2**	**2**	**K00089**	**2HDJ**	**SSF51430**	**PF00248**
Oxidorredutase	**1.3.1.20**	***trans-1,2-dihydrobenzene-1,2-diol dehydrogenase***	***Metabolism of xenobiotics by cytochrome P450 Pentose and glucuronate interconversions Steroid hormone biosynthesis***	**AKR1C2**	**2**	**K00089**	**2HDJ**	**SSF51430**	**PF00248**
Oxidorredutase	**1.3.1.20**	***trans-1,2-dihydrobenzene-1,2-diol dehydrogenase***	***Metabolism of xenobiotics by cytochrome P450 Pentose and glucuronate interconversions Steroid hormone biosynthesis***	**AKR1C1**	**2**	**K00212**	**1J96**	**SSF51430**	**PF00248**

Classe	EC	Atividade	Via Metabólica	Gene	AnEnPi	KO	PDB	SUPERFAMILY	Pfam
Transferase	**2.4.1.22**	***lactose synthase***	***Galactose metabolism Glycosaminoglycan biosynthesis - keratan sulfate Glycosphingolipid biosynthesis - lacto and neolacto series Mannose type O-glycan biosynthesis N-Glycan biosynthesis Other types of O-glycan biosynthesis***	**B4GALT2**	**1**	**K07967**	**ND**	**SSF53448**	**PF02709 PF13733**
Transferase	**2.4.1.22**	***lactose synthase***	***Galactose metabolism Glycosaminoglycan biosynthesis - keratan sulfate Glycosphingolipid biosynthesis - lacto and neolacto series Mannose type O-glycan biosynthesis N-Glycan biosynthesis Other types of O-glycan biosynthesis***	**B4GALT1**	**1**	**K07966**	**2AH9**	**SSF53448**	**PF02709 PF13733**

Classe	EC	Atividade	Via Metabólica	Gene	AnEnPi	KO	PDB	SUPERFAMILY	Pfam
Transferase	**2.4.1.22**	***lactose synthase***	***Galactose metabolism Glycosaminoglycan biosynthesis - keratan sulfate Glycosphingolipid biosynthesis - lacto and neolacto series Mannose type O-glycan biosynthesis N-Glycan biosynthesis Other types of O-glycan biosynthesis***	**LALBA**	**2**	**K00704**	**3B0O**	**SSF53955**	**PF00062**
Transferase	**2.4.2.31**	***NAD+—protein-arginine ADP-ribosyltransferase***	***Central carbon metabolism in cancer***	**SIRT6**	**1**	**K11416**	**3K35**	**SSF52467**	**PF02146**
Transferase	**2.4.2.31**	***NAD+—protein-arginine ADP-ribosyltransferase***	***Central carbon metabolism in cancer***	**ART1**	**2**	**K06716**	**ND**	**SSF56399**	**PF01129**
Transferase	**2.4.2.31**	***NAD+—protein-arginine ADP-ribosyltransferase***	***Central carbon metabolism in cancer***	**ART3**	**2**	**K00775**	**ND**	**SSF56399**	**PF01129**
Transferase	**2.4.2.31**	***NAD+—protein-arginine ADP-ribosyltransferase***	***Central carbon metabolism in cancer***	**ART4**	**2**	**K06717**	**ND**	**SSF56399**	**PF01129**
Transferase	**2.4.2.31**	***NAD+—protein-arginine ADP-ribosyltransferase***	***Central carbon metabolism in cancer***	**ART5**	**2**	**K19977**	**ND**	**SSF56399**	**PF01129**
Transferase	**2.7.1.67**	***1-phosphatidylinositol 4-kinase***	***Inositol phosphate metabolism Phosphatidylinositol signaling system***	**PI4KA**	**1**	**K00888**	**ND**	**SSF48371 SSF56112**	**PF00454 PF00613**
Transferase	**2.7.1.67**	***1-phosphatidylinositol 4-kinase***	***Inositol phosphate metabolism Phosphatidylinositol signaling system***	**PI4KB**	**1**	**K19801**	**4WAE**	**SSF56112**	**PF00454**

Classe	EC	Atividade	Via Metabólica	Gene	AnEnPi	KO	PDB	SUPERFAMILY	Pfam
Transferase	**2.7.1.67**	***1-phosphatidylinositol 4-kinase***	***Inositol phosphate metabolism Phosphatidylinositol signaling system***	**PI4K2A**	**2**	**K13711**	**4HND**	**SSF56399**	**PF00454**
Transferase	**2.7.1.67**	***1-phosphatidylinositol 4-kinase***	***Inositol phosphate metabolism Phosphatidylinositol signaling system***	**PI4K2B**	**2**	**K13711**	**4WTV**	**ND**	**PF00454**
Transferase	**2.7.4.21**	***inositol-hexakisphosphate kinase***	***Phosphatidylinositol signaling system***	**PPIP5K2**	**1**	**K13024**	**3T9A**	**SSF53254 SSF56059**	**PF00328**
Transferase	**2.7.4.21**	***inositol-hexakisphosphate kinase***	***Phosphatidylinositol signaling system***	**PPIP5K1**	**1**	**K13024**	**ND**	**SSF53254 SSF56059**	**PF00328**
Transferase	**2.7.4.21**	***inositol-hexakisphosphate kinase***	***Phosphatidylinositol signaling system***	**IP6K1**	**2**	**K07756**	**ND**	**SSF56104**	**PF03770**
Transferase	**2.7.4.21**	***inositol-hexakisphosphate kinase***	***Phosphatidylinositol signaling system***	**IP6K3**	**2**	**K07756**	**ND**	**SSF56104**	**PF03770**
Transferase	**2.7.4.21**	***inositol-hexakisphosphate kinase***	***Phosphatidylinositol signaling system***	**IP6K2**	**2**	**K07756**	**ND**	**SSF56104**	**PF03770**
Hidrolase	**3.1.1.29**	***aminoacyl-tRNA hydrolase***	***ND***	**PTRH2**	**1**	**K04794**	**1Q7S**	**SSF102462**	**PF01981**
Hidrolase	**3.1.1.29**	***aminoacyl-tRNA hydrolase***	***ND***	**PTRH1**	**2**	**K01056**	**ND**	**SSF53178**	**PF01195**
Hidrolase	**3.1.1.29**	***aminoacyl-tRNA hydrolase***	***ND***	**ICT1**	**3**	**K15033**	**ND**	**39326**	**PF00472**

Classe	EC	Atividade	Via Metabólica	Gene	AnEnPi	KO	PDB	SUPERFAMILY	Pfam
Hidrolase	**3.1.1.3**	***triacylglycerol lipase***	***Fat digestion and absorption Glycerolipid metabolism Pancreatic secretion Regulation of lipolysis in adipocytes Steroid biosynthesis Vitamin digestion and absorption***	**AADAC**	**3**	**K13616**	**ND**	**SSF53474**	**PF07859**
Hidrolase	**3.1.1.3**	***triacylglycerol lipase***	***Fat digestion and absorption Glycerolipid metabolism Pancreatic secretion Regulation of lipolysis in adipocytes Steroid biosynthesis Vitamin digestion and absorption***	**CEL**	**3**	**K12298**	**ND**	**SSF53474**	**PF00135**
Hidrolase	**3.1.1.3**	***triacylglycerol lipase***	***Fat digestion and absorption Glycerolipid metabolism Pancreatic secretion Regulation of lipolysis in adipocytes Steroid biosynthesis Vitamin digestion and absorption***	**LIPC**	**4**	**K01046**	**ND**	**SSF49723 SSF53474**	**PF00151 PF01477**

Classe	EC	Atividade	Via Metabólica	Gene	AnEnPi	KO	PDB	SUPERFAMILY	Pfam
Hidrolase	**3.1.1.3**	***triacylglycerol lipase***	***Fat digestion and absorption Glycerolipid metabolism Pancreatic secretion Regulation of lipolysis in adipocytes Steroid biosynthesis Vitamin digestion and absorption***	**PNLIP**	**4**	**K14073**	**1LPB**	**SSF49723 SSF53474**	**PF00151 PF01477**
Hidrolase	**3.1.1.3**	***triacylglycerol lipase***	***Fat digestion and absorption Glycerolipid metabolism Pancreatic secretion Regulation of lipolysis in adipocytes Steroid biosynthesis Vitamin digestion and absorption***	**PNLIPRP1**	**4**	**K14074**	**2PPL**	**SSF49723 SSF53474**	**PF00151 PF01477**
Hidrolase	**3.1.1.3**	***triacylglycerol lipase***	***Fat digestion and absorption Glycerolipid metabolism Pancreatic secretion Regulation of lipolysis in adipocytes Steroid biosynthesis Vitamin digestion and absorption***	**PNLIPRP3**	**4**	**K14076**	**ND**	**SSF49723 SSF53474**	**PF00151 PF01477**

Classe	EC	Atividade	Via Metabólica	Gene	AnEnPi	KO	PDB	SUPERFAMILY	Pfam
Hidrolase	**3.1.1.3**	***triacylglycerol lipase***	***Fat digestion and absorption Glycerolipid metabolism Pancreatic secretion Regulation of lipolysis in adipocytes Steroid biosynthesis Vitamin digestion and absorption***	**LIPG**	**4**	**K01046**	**ND**	**SSF49723 SSF53474**	**PF00151 PF01477**
Hidrolase	**3.1.1.3**	***triacylglycerol lipase***	***Fat digestion and absorption Glycerolipid metabolism Pancreatic secretion Regulation of lipolysis in adipocytes Steroid biosynthesis Vitamin digestion and absorption***	**PNLIPRP2**	**4**	**K14075**	**2OXE**	**SSF49723 SSF53474**	**PF00151 PF01477**
Hidrolase	**3.1.1.3**	***triacylglycerol lipase***	***Fat digestion and absorption Glycerolipid metabolism Pancreatic secretion Regulation of lipolysis in adipocytes Steroid biosynthesis Vitamin digestion and absorption***	**LIPF**	**9**	**K14452**	**1HLG**	**SSF53474**	**PF00561 PF04083**

Classe	EC	Atividade	Via Metabólica	Gene	AnEnPi	KO	PDB	SUPERFAMILY	Pfam
Hidrolase	**3.1.1.3**	***triacylglycerol lipase***	***Fat digestion and absorption Glycerolipid metabolism Pancreatic secretion Regulation of lipolysis in adipocytes Steroid biosynthesis Vitamin digestion and absorption***	**PNPLA3**	**10**	**K13534**	**ND**	**SSF52151**	**PF01734**
Hidrolase	**3.1.2.2**	***palmitoyl-CoA hydrolase***	***Bile secretion Biosynthesis of unsaturated fatty acids Fatty acid elongation Ovarian steroidogenesis Peroxisome Primary bile acid biosynthesis Taurine and hypotaurine metabolism***	**ACOT2**	**1**	**K01068**	**3HLK**	**SSF53474**	**PF04775 PF08840**
Hidrolase	**3.1.2.2**	***palmitoyl-CoA hydrolase***	***Bile secretion Biosynthesis of unsaturated fatty acids Fatty acid elongation Ovarian steroidogenesis Peroxisome Primary bile acid biosynthesis Taurine and hypotaurine metabolism***	**BAAT**	**1**	**K00659**	**ND**	**SSF53474**	**PF04775 PF08840**

Classe	EC	Atividade	Via Metabólica	Gene	AnEnPi	KO	PDB	SUPERFAMILY	Pfam
Hidrolase	3.1.2.2	*palmitoyl-CoA hydrolase*	*Bile secretion* *Biosynthesis of unsaturated fatty acids* *Fatty acid elongation* *Ovarian steroidogenesis* *Peroxisome* *Primary bile acid biosynthesis* *Taurine and hypotaurine metabolism*	ACOT1	1	K01068	ND	SSF53474	PF04775 PF08840
Hidrolase	3.1.2.2	*palmitoyl-CoA hydrolase*	*Bile secretion* *Biosynthesis of unsaturated fatty acids* *Fatty acid elongation* *Ovarian steroidogenesis* *Peroxisome* *Primary bile acid biosynthesis* *Taurine and hypotaurine metabolism*	ACOT4	1	K01068	3K2I	SSF53474	PF04775 PF08840
Hidrolase	3.1.2.2	*palmitoyl-CoA hydrolase*	*Bile secretion* *Biosynthesis of unsaturated fatty acids* *Fatty acid elongation* *Ovarian steroidogenesis* *Peroxisome* *Primary bile acid biosynthesis* *Taurine and hypotaurine metabolism*	ACOT7	2	K17360	2QQ2	SSF54637	PF03061

Classe	EC	Atividade	Via Metabólica	Gene	AnEnPi	KO	PDB	SUPERFAMILY	Pfam
Hidrolase	**3.1.3.2**	***acid phosphatase***	***Adherens junction Lysosome Osteoclast differentiation Rheumatoid arthritis Thiamine metabolism***	**ACP5**	**1**	**K14379**	**1WAR**	**SSF56300**	**PF00149**
Hidrolase	**3.1.3.2**	***acid phosphatase***	***Adherens junction Lysosome Osteoclast differentiation Rheumatoid arthritis Thiamine metabolism***	**ACP2**	**2**	**K14410**	**ND**	**SSF53254**	**PF00328**
Hidrolase	**3.1.3.2**	***acid phosphatase***	***Adherens junction Lysosome Osteoclast differentiation Rheumatoid arthritis Thiamine metabolism***	**ACPP**	**2**	**K19283**	**1CVI**	**SSF53254**	**PF00328**
Hidrolase	**3.1.3.2**	***acid phosphatase***	***Adherens junction Lysosome Osteoclast differentiation Rheumatoid arthritis Thiamine metabolism***	**ACPT**	**2**	**K19284**	**ND**	**SSF53254**	**PF00328**
Hidrolase	**3.1.3.2**	***acid phosphatase***	***Adherens junction Lysosome Osteoclast differentiation Rheumatoid arthritis Thiamine metabolism***	**ACP6**	**2**	**K14395**	**4JOB**	**SSF53254**	**PF00328**

Classe	EC	Atividade	Via Metabólica	Gene	AnEnPi	KO	PDB	SUPERFAMILY	Pfam
Hidrolase	**3.1.3.2**	***acid phosphatase***	***Adherens junction*** ***Lysosome*** ***Osteoclast differentiation*** ***Rheumatoid arthritis*** ***Thiamine metabolism***	**ACP1**	**5**	**K14394**	**5PNT**	**SSF52788**	**PF01451**
Hidrolase	**3.1.3.5**	***5'-nucleotidase***	***Nicotinate and nicotinamide metabolism*** ***Purine metabolism*** ***Pyrimidine metabolism***	**NT5C1B**	**2**	**K01081**	**ND**	**ND**	**PF06189**
Hidrolase	**3.1.3.5**	***5'-nucleotidase***	***Nicotinate and nicotinamide metabolism*** ***Purine metabolism*** ***Pyrimidine metabolism***	**NT5C1A**	**2**	**K01081**	**ND**	**ND**	**PF06189**
Hidrolase	**3.1.3.5**	***5'-nucleotidase***	***Nicotinate and nicotinamide metabolism*** ***Purine metabolism*** ***Pyrimidine metabolism***	**NT5E**	**3**	**K19970**	**4H2G**	**SSF55816** **SSF56300**	**PF00149** **PF02872**
Hidrolase	**3.1.3.5**	***5'-nucleotidase***	***Nicotinate and nicotinamide metabolism*** ***Purine metabolism*** ***Pyrimidine metabolism***	**NT5C**	**5**	**K01081**	**4L57**	**SSF56784**	**PF06941**
Hidrolase	**3.1.3.5**	***5'-nucleotidase***	***Nicotinate and nicotinamide metabolism*** ***Purine metabolism*** ***Pyrimidine metabolism***	**NT5M**	**5**	**K01081**	**4MUM**	**SSF56784**	**PF06941**
Hidrolase	**3.1.3.5**	***5'-nucleotidase***	***Nicotinate and nicotinamide metabolism*** ***Purine metabolism*** ***Pyrimidine metabolism***	**NT5C3A**	**7**	**K01081**	**2CN1**	**SSF56784**	**PF05822**

Classe	EC	Atividade	Via Metabólica	Gene	AnEnPi	KO	PDB	SUPERFAMILY	Pfam
Hidrolase	3.1.3.5	*5'-nucleotidase*	*Nicotinate and nicotinamide metabolism* *Purine metabolism* *Pyrimidine metabolism*	NT5C2	9	K01081	2XCW	SSF56784	PF05761
Hidrolase	3.1.4.12	*sphingomyelin phosphodiesterase*	*Lysosome* *Sphingolipid metabolism* *Sphingolipid signaling pathway*	SMPD2	1	K12351	ND	SSF56219	PF03372
Hidrolase	3.1.4.12	*sphingomyelin phosphodiesterase*	*Lysosome* *Sphingolipid metabolism* *Sphingolipid signaling pathway*	SMPD3	1	K12352	ND	SSF56219	PF03372
Hidrolase	3.1.4.12	*sphingomyelin phosphodiesterase*	*Lysosome* *Sphingolipid metabolism* *Sphingolipid signaling pathway*	SMPD1	2	K12350	5I81	SSF47862 SSF56300	PF00149
Hidrolase	3.1.4.12	*sphingomyelin phosphodiesterase*	*Lysosome* *Sphingolipid metabolism* *Sphingolipid signaling pathway*	SMPD4	3	K12353	ND	ND	PF14724
Hidrolase	3.1.4.12	*sphingomyelin phosphodiesterase*	*Lysosome* *Sphingolipid metabolism* *Sphingolipid signaling pathway*	ENPP7	4	K12354	5UDY	SSF53649	PF01663

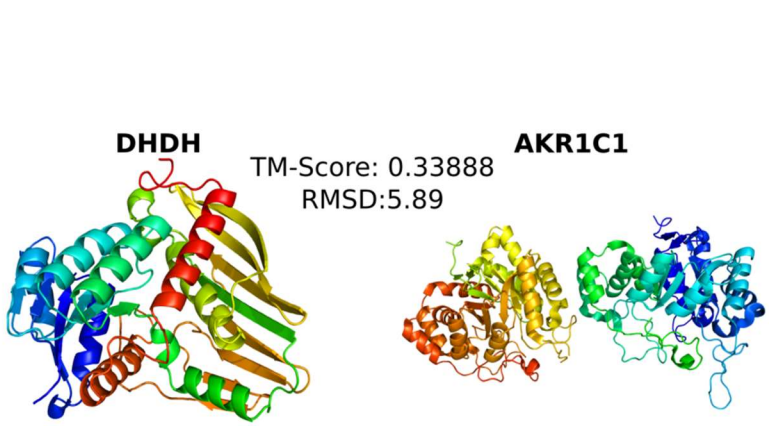
Classe	EC	Atividade	Via Metabólica	Gene	AnEnPi	KO	PDB	SUPERFAMILY	Pfam
Liase	**4.2.99.18**	***DNA-(apurinic or apyrimidinic site) lyase***	***Base excision repair***	**NTHL1**	**1**	**K10773**	**ND**	**SSF48150**	**PF00633 PF00730**
Liase	**4.2.99.18**	***DNA-(apurinic or apyrimidinic site) lyase***	***Base excision repair***	**OGG1**	**1**	**K03660**	**1KO9**	**SSF48150 SSF55945**	**PF00730 PF07934**
Liase	**4.2.99.18**	***DNA-(apurinic or apyrimidinic site) lyase***	***Base excision repair***	**NEIL2**	**1**	**K10568**	**1VZP**	**SSF46946**	**PF01149 PF06831**
Liase	**4.2.99.18**	***DNA-(apurinic or apyrimidinic site) lyase***	***Base excision repair***	**NEIL1**	**1**	**K10567**	**1TDH**	**SSF46946 SSF57716 SSF81624**	**PF01149 PF06831 PF09292**
Liase	**4.2.99.18**	***DNA-(apurinic or apyrimidinic site) lyase***	***Base excision repair***	**APEX1**	**1**	**K10771**	**2O3H**	**SSF56219**	**PF03372**
Liase	**4.2.99.18**	***DNA-(apurinic or apyrimidinic site) lyase***	***Base excision repair***	**APEX2**	**1**	**K10772**	**ND**	**SSF56219**	**PF03372 PF06839**
Liase	**4.2.99.18**	***DNA-(apurinic or apyrimidinic site) lyase***	***Base excision repair***	**APLF**	**6**	**K13295**	**2KUO**	**SSF49879**	**PF10283**

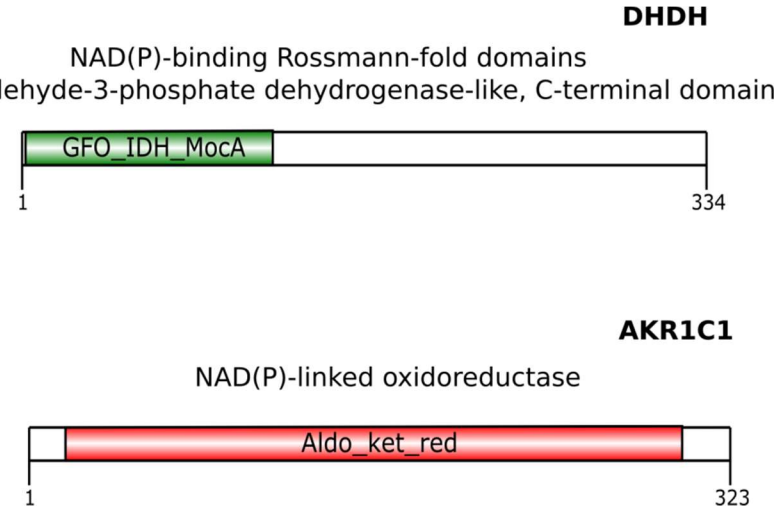
Classe	EC	Atividade	Via Metabólica	Gene	AnEnPi	KO	PDB	SUPERFAMILY	Pfam
Isomerase	**5.3.99.2**	***prostaglandin-D synthase***	***Arachidonic acid metabolism Chemical carcinogenesis Drug metabolism - cytochrome P450 Glutathione metabolism Metabolism of xenobiotics by cytochrome P450***	**PTGDS**	**1**	**K01830**	**2WWP**	**SSF50814**	**PF00061**
Isomerase	**5.3.99.2**	***prostaglandin-D synthase***	***Arachidonic acid metabolism Chemical carcinogenesis Drug metabolism - cytochrome P450 Glutathione metabolism Metabolism of xenobiotics by cytochrome P450***	**HPGDS**	**2**	**K04097**	**1IYI**	**SSF47616 SSF52833**	**PF00043 PF02798**
Isomerase	**5.3.99.3**	***prostaglandin-E synthase***	***Arachidonic acid metabolism***	**PTGES2**	**1**	**K05309**	**ND**	**SSF47616 SSF52833**	**PF13417 PF14497**
Isomerase	**5.3.99.3**	***prostaglandin-E synthase***	***Arachidonic acid metabolism***	**PTGES**	**2**	**K15729**	**4AL0**	**SSF161084**	**PF01124**

Anexo II: Comparação de estrutura tridimensional (3D) e arquitetura de domínios de todas as enzimas análogas intragenômicas humanas identificadas em nosso estudo com informações estruturais e de domínio (estruturas 3D resolvidas experimentalmente, obtidas no PDB, ou modelos estruturais para sequências sem informação 3D, e classificação Pfam, respectivamente). Cada figura representa uma atividade enzimática diferente, na qual são apresentadas as distintas composição de domínios e classificação em superfamília de seqüências representativas, bem como suas estruturas 3D e os resultados de alinhamentos estruturais entre pares de sequências enzimáticas. Os valores de TM-score e RMSD são mostrados para cada par de estruturas 3D comparadas. As cores das estruturas estão representadas em *Rainbow*, com faixas de cores variando do azul (N-terminal) até o vermelho (C-terminal).Os diagramas de arquitetura de domínio foram desenhados com a ferramenta IBS (109).

EC:1.3.1.20

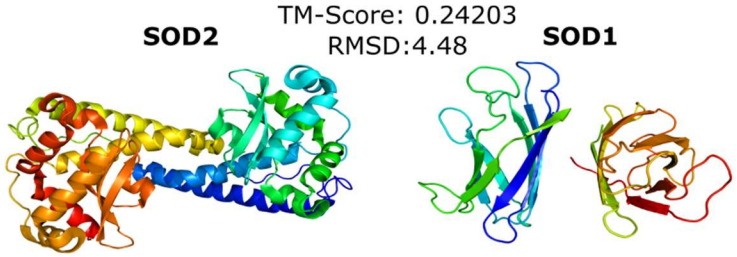
Trans-1,2-Dihydrobenzene-1,2-Diol Dehydrogenase

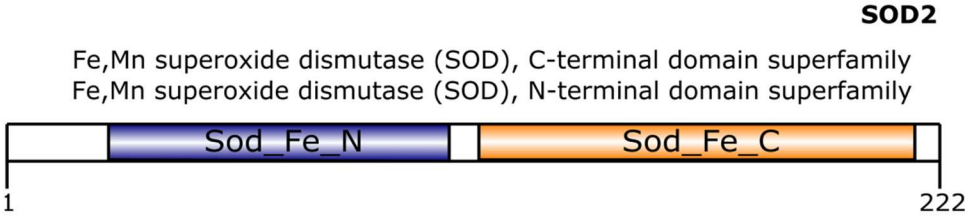


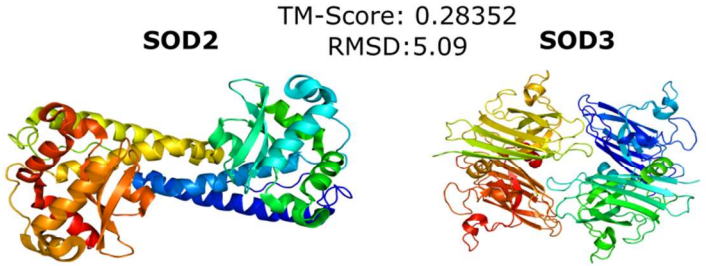


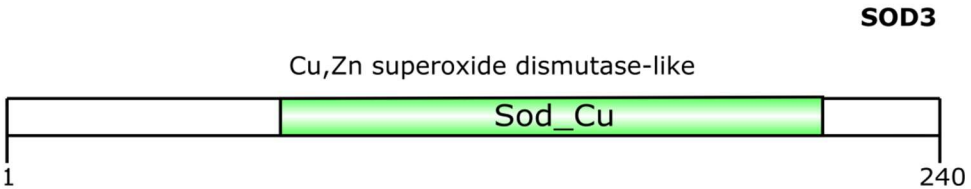
EC:1.15.1.1

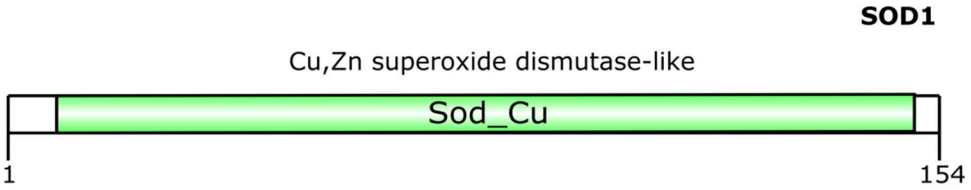
Superoxide Dismutase

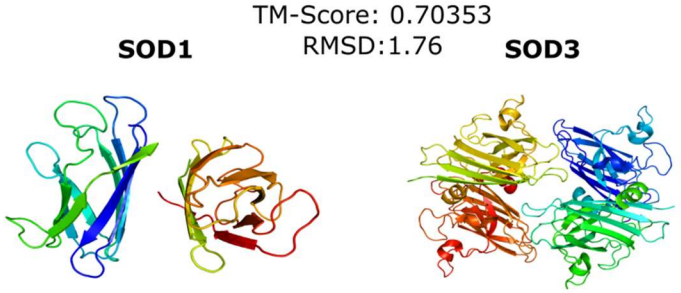






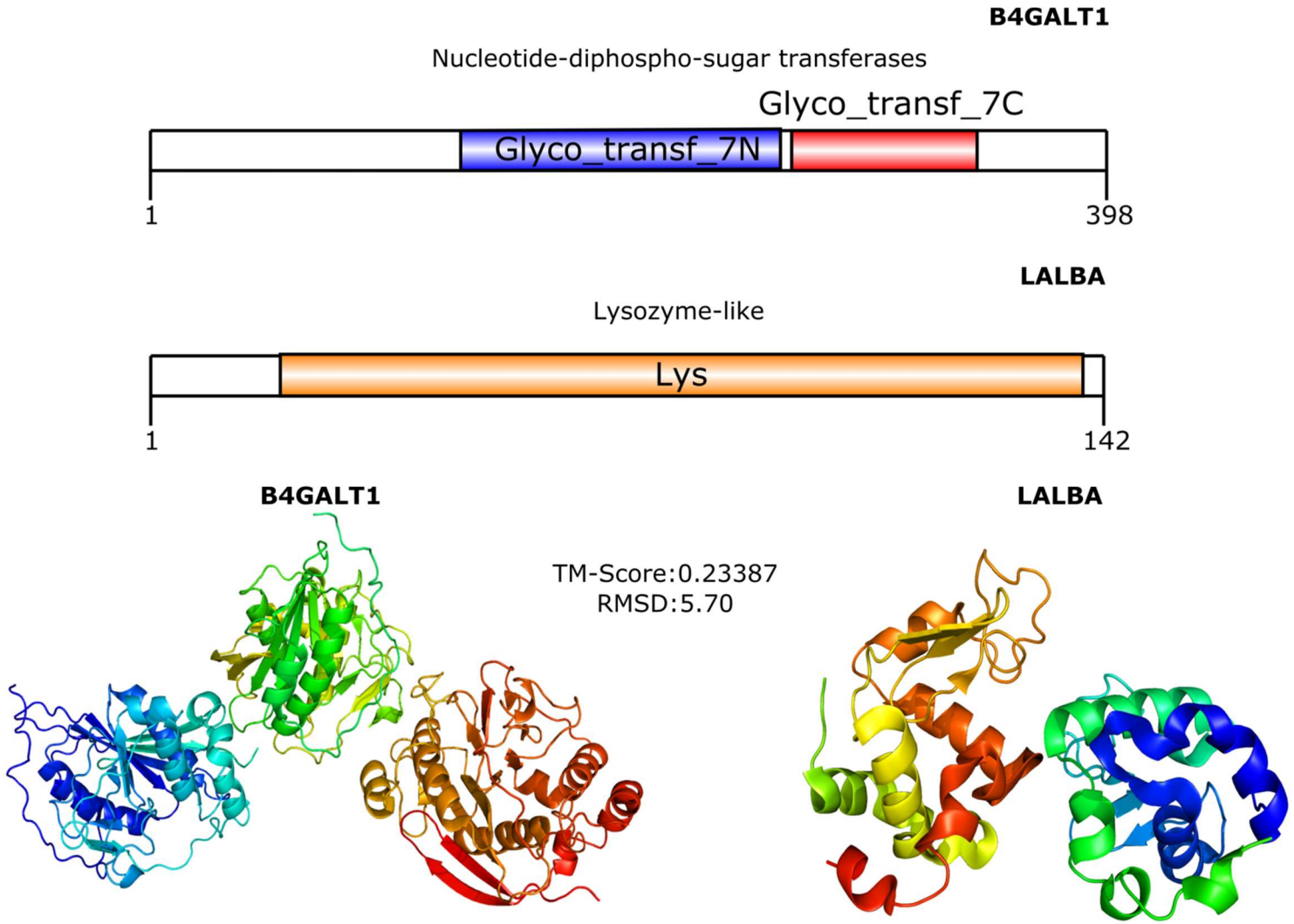






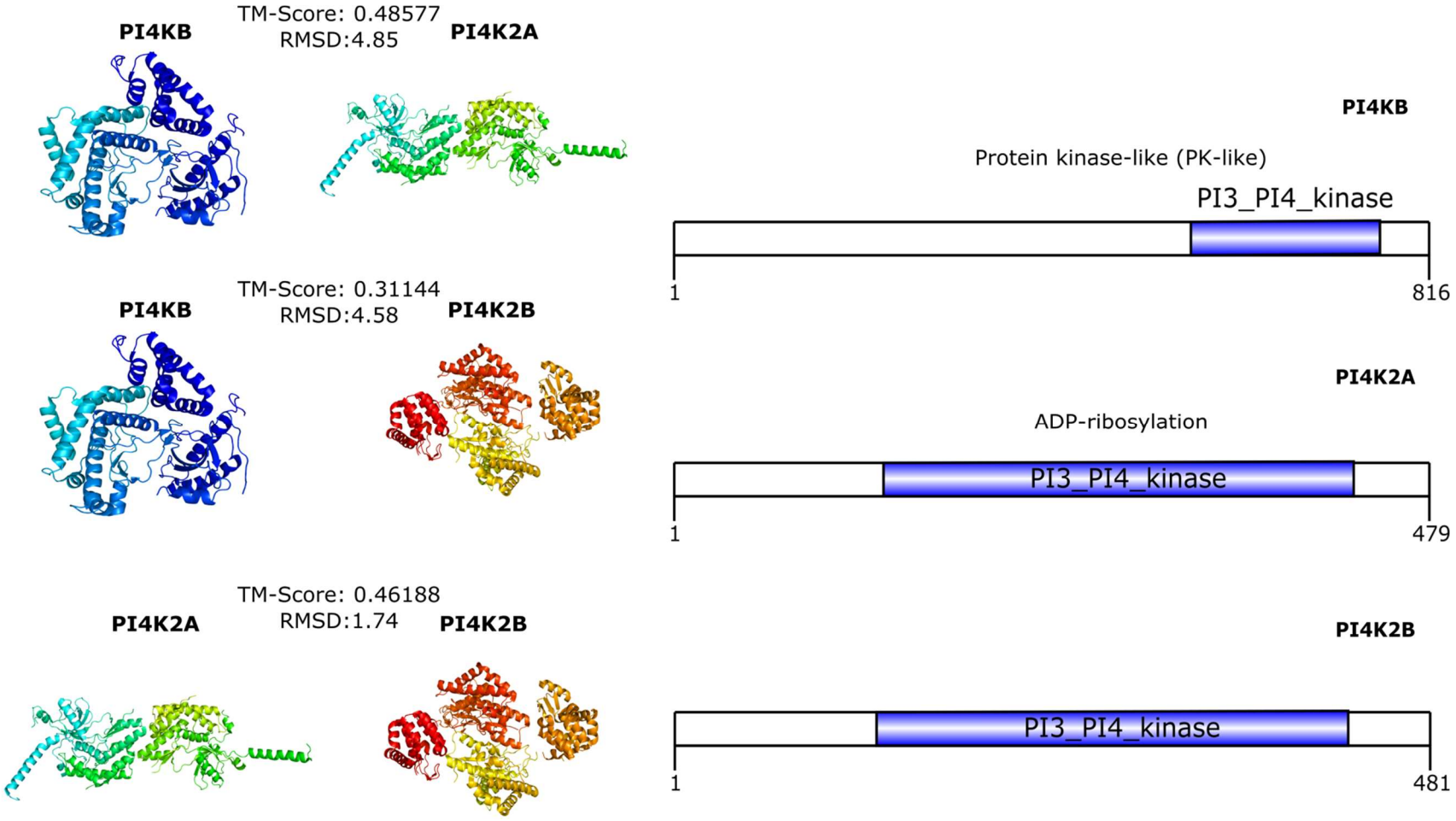
EC:2.4.1.22

Lactose Synthase



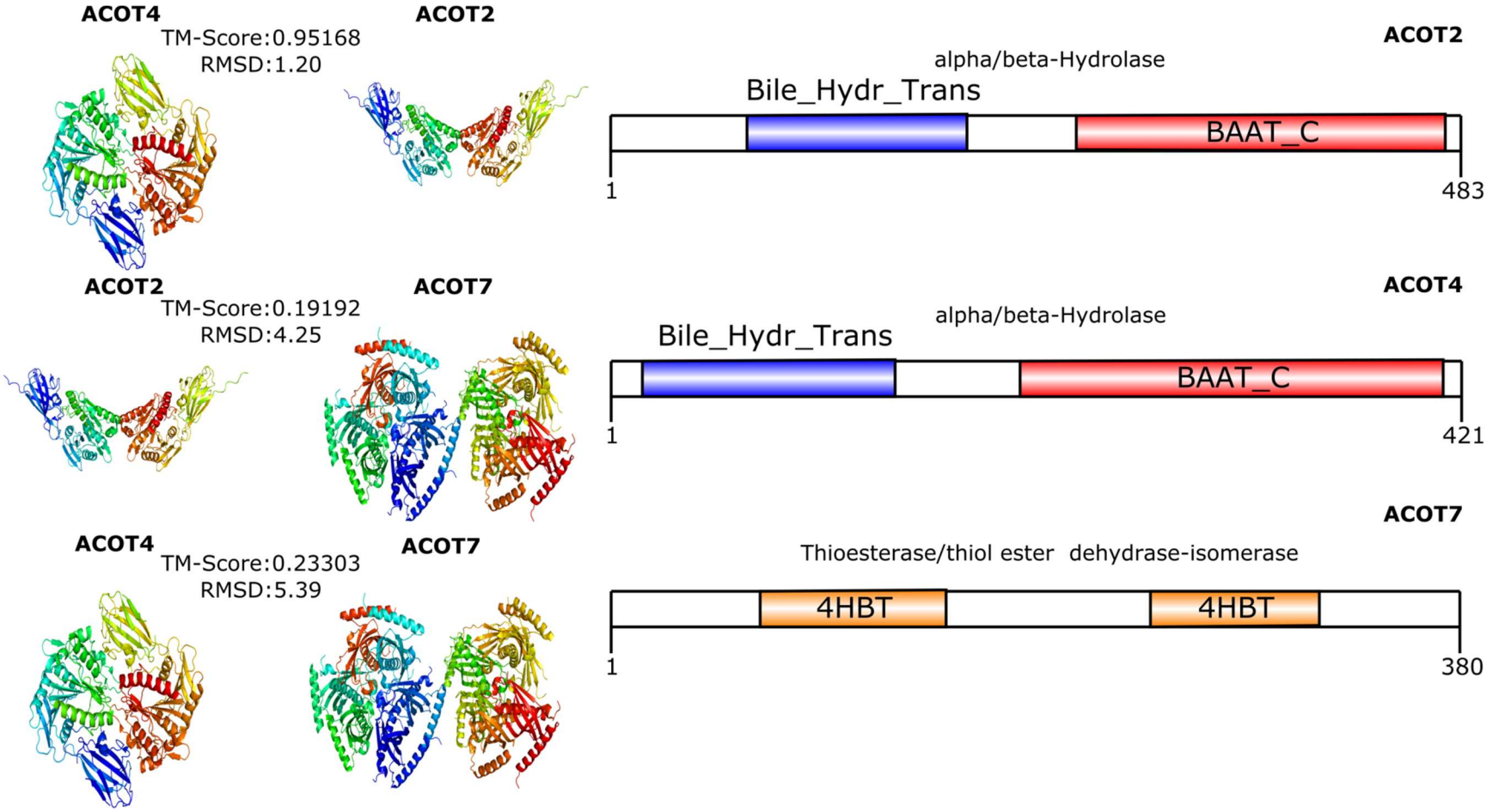
EC:2.7.1.67

1-Phosphatidylinositol 4-Kinase



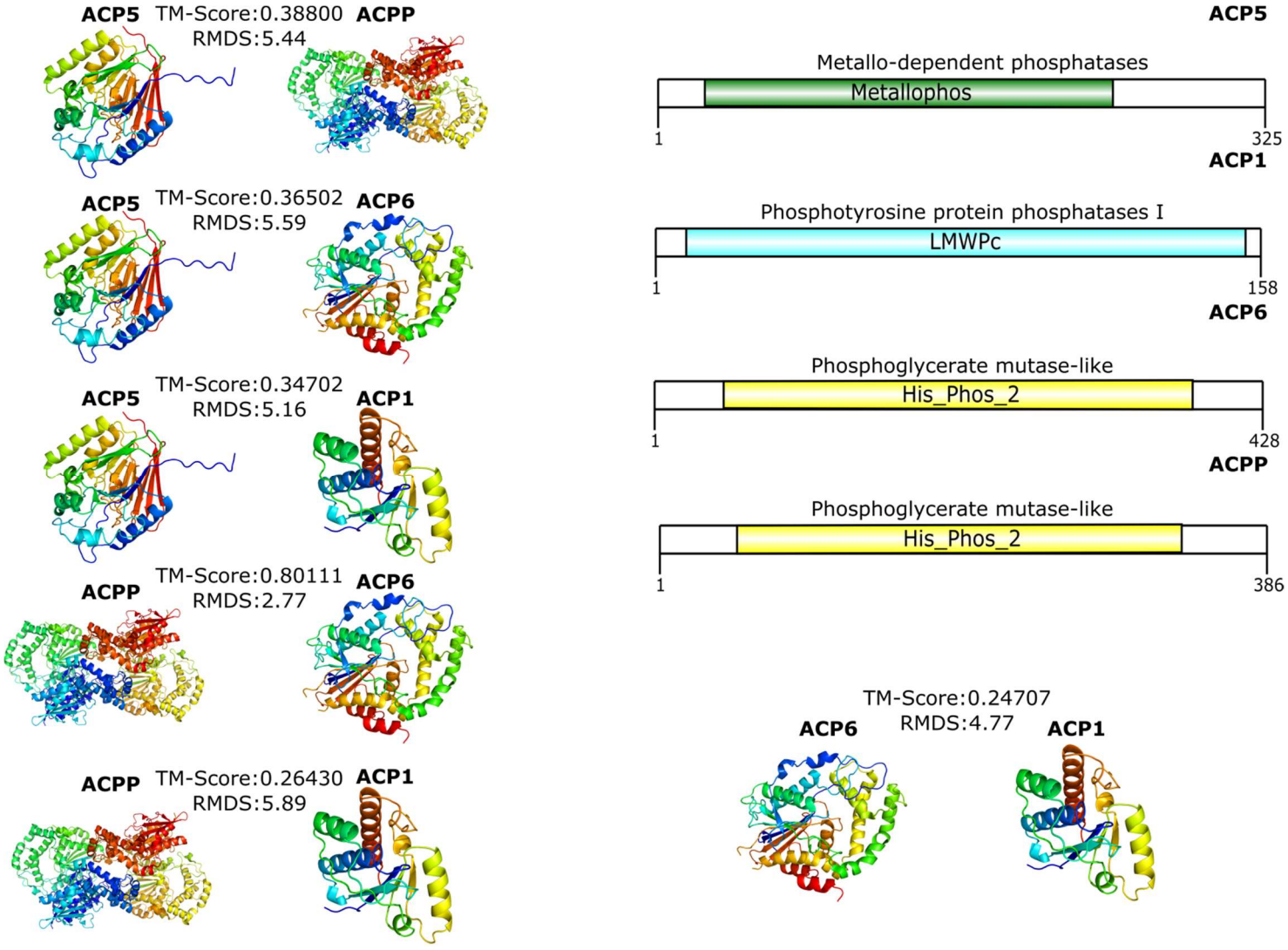
EC:3.1.2.2

Palmitoyl-CoA Hydrolase



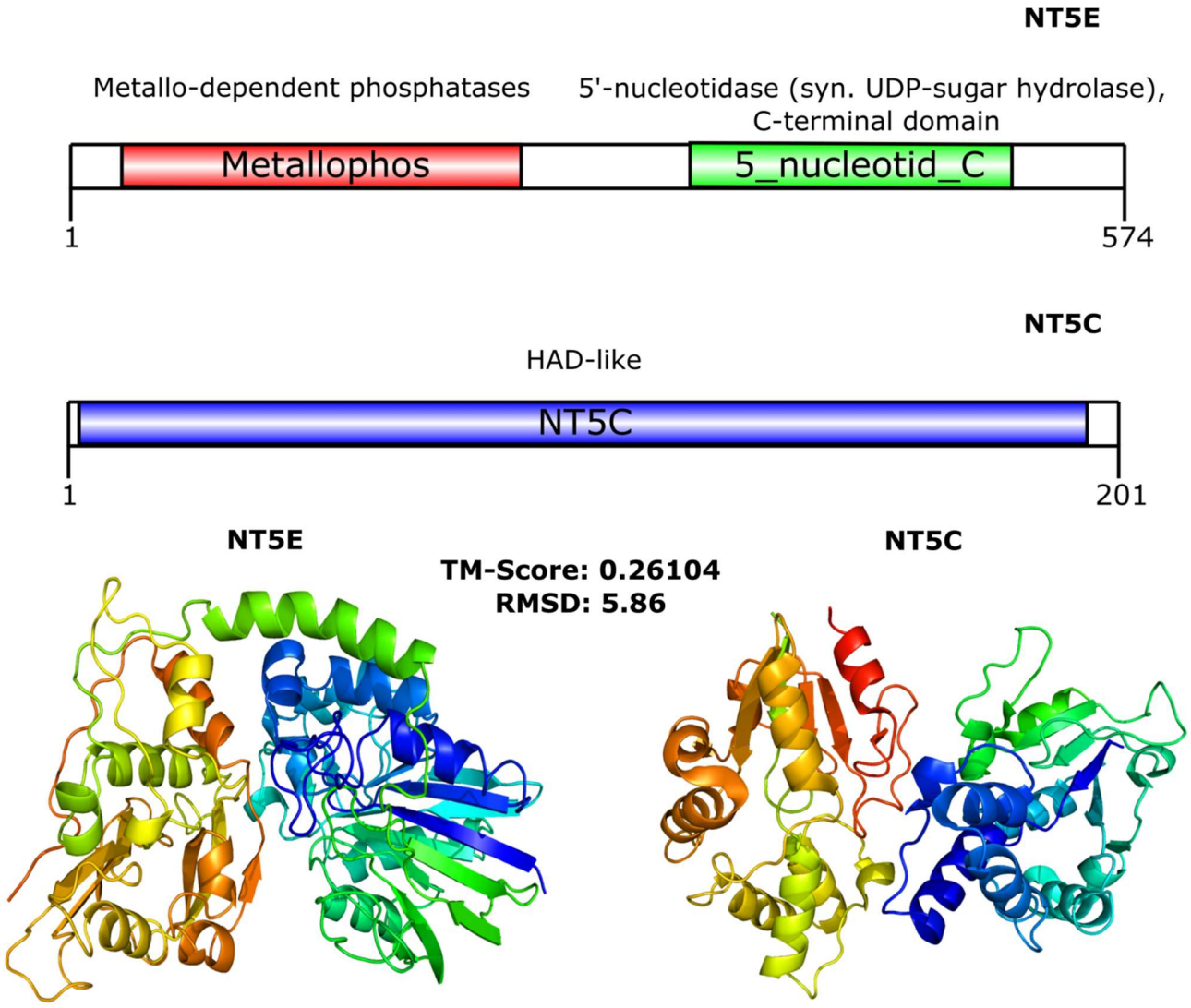
EC:3.1.3.2

Acid Phosphatase



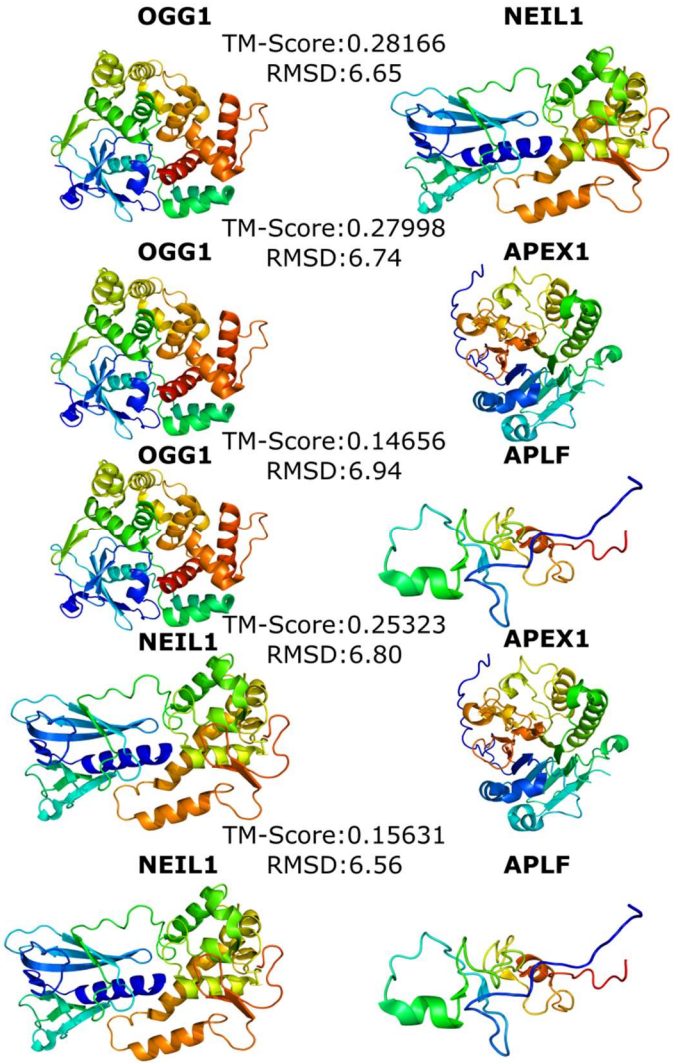
EC:3.1.3.5

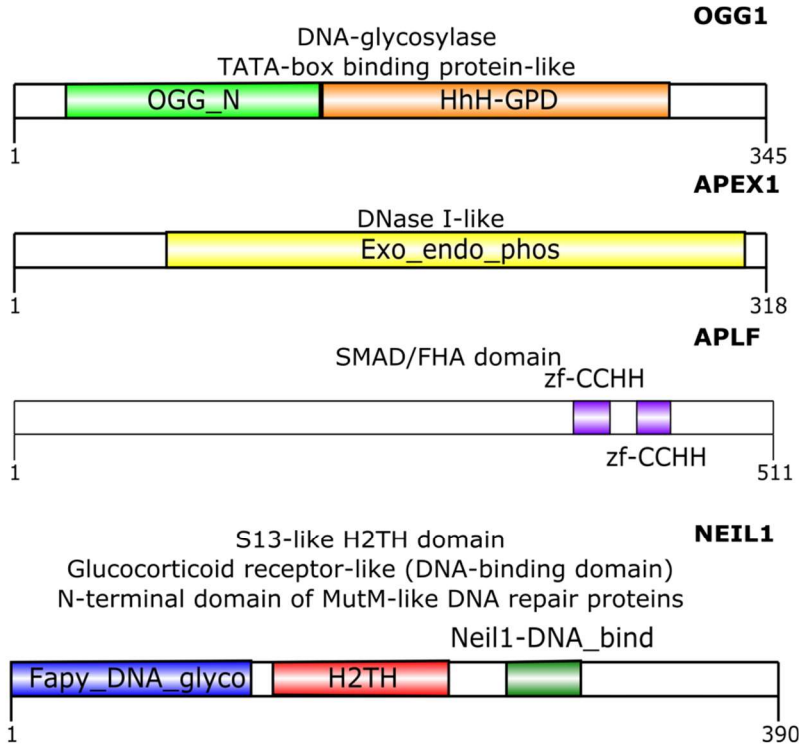
5'-nucleotidase

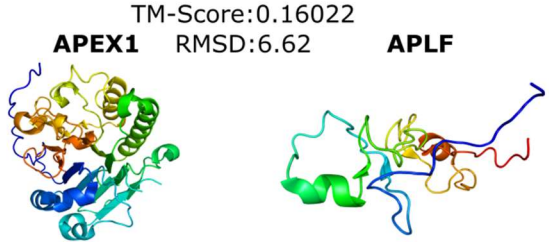


EC:4.2.99.18

DNA-(apurinic or apyrimidinic site) lyase

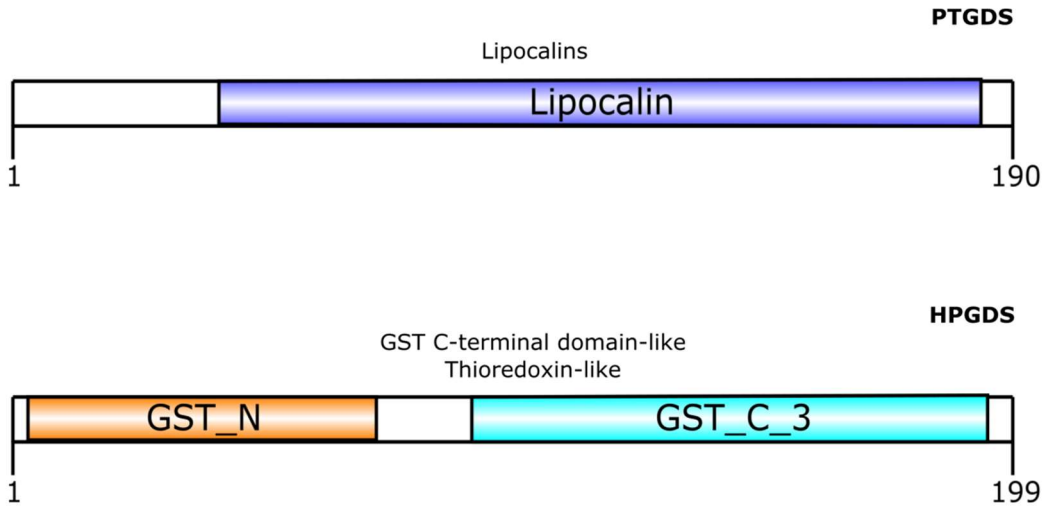


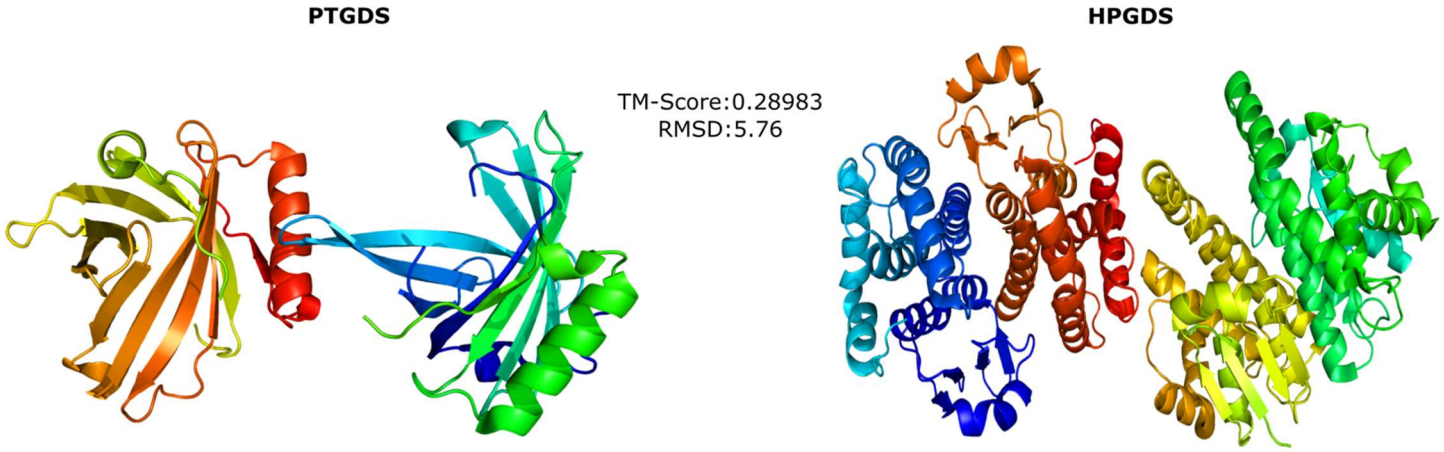




EC:5.3.99.2

Prostaglandin-D Synthase





Anexo III: Artigo "*Functional Analogy in Human Metabolism: Enzymes with Different Biological Roles or Functional Redundancy?*" publicado na revista "*Genome Biology and Evolution*" em 06 de julho de 2017, contendo parte dos dados apresentados nessa tese.

Functional Analogy in Human Metabolism: Enzymes with Different Biological Roles or Functional Redundancy?

Rafael Mina Piergiorge[1], Antonio Basílio de Miranda[2], Ana Carolina Guimarães[1,*], and Marcos Catanho[1]

[1]Laboratório de Genômica Funcional e Bioinformática, Fiocruz, Instituto Oswaldo Cruz, Manguinhos, Rio de Janeiro, Brazil

[2]Laboratório de Biologia Computacional e Sistemas, Fiocruz, Instituto Oswaldo Cruz, Manguinhos, Rio de Janeiro, Brazil

*Corresponding author: E-mail: carolg@fiocruz.br.

Accepted: July 4, 2017

Abstract

Since enzymes catalyze almost all chemical reactions that occur in living organisms, it is crucial that genes encoding such activities are correctly identified and functionally characterized. Several studies suggest that the fraction of enzymatic activities in which multiple events of independent origin have taken place during evolution is substantial. However, this topic is still poorly explored, and a comprehensive investigation of the occurrence, distribution, and implications of these events has not been done so far. Fundamental questions, such as how analogous enzymes originate, why so many events of independent origin have apparently occurred during evolution, and what are the reasons for the coexistence in the same organism of distinct enzymatic forms catalyzing the same reaction, remain unanswered. Also, several isofunctional enzymes are still not recognized as nonhomologous, even with substantial evidence indicating different evolutionary histories. In this work, we begin to investigate the biological significance of the cooccurrence of nonhomologous isofunctional enzymes in human metabolism, characterizing functional analogous enzymes identified in metabolic pathways annotated in the human genome. Our hypothesis is that the coexistence of multiple enzymatic forms might not be interpreted as functional redundancy. Instead, these enzymatic forms may be implicated in distinct (and probably relevant) biological roles.

Key words: enzymatic activity, convergent evolution, *H. sapiens*.

De Novo Origin of Enzymatic Activities and Functional Analogy in Human Metabolism

Enzymes have their biological activities defined by the type of chemical transformation carried out and by the mechanism through which this reaction is executed. The chemical transformations accomplished by enzymes are classified using the recommendations of the Nomenclature Committee of the International Union of Biochemistry (http://www.chem.qmul.ac.uk/iubmb/enzyme/). Based on the reaction catalyzed by the enzyme an Enzyme Commission (EC) number is assigned. According to this hierarchical classification, each enzyme receives a 4-digit number: the first digit describes the general chemical reaction catalyzed by the enzyme (the enzyme class); the two subsequent numbers have different meanings depending on the class of the enzyme; the fourth digit describes the specificity of the reaction, defining the specific substrate/product or cofactors used (McDonald and Tipton 2014).

In silico comparisons of metabolic pathways predicted from completely sequenced genomes of a variety of prokaryotic and eukaryotic species revealed incomplete or even absent pathways in several organisms (Cordwell 1999; Galperin and Koonin 1999; Huynen et al. 1999; Morett et al. 2003; Peregrin-Alvarez et al. 2003; Hanson et al. 2010). In some of these cases, the "missing" enzymes were replaced by functional equivalent molecules, able to catalyze the same reaction but exhibiting virtually no similarity in their amino acid chains, thus escaping identification by methods based on sequence similarity. These nonhomologous isofunctional molecules, known as analogous enzymes, arise from independent evolutionary events, converging for the same biological function, and may be associated with both related or unrelated phylogenetic lineages and/or possess different catalytic mechanisms, as well as distinct fold topologies and three-dimensional (3D) structures (Cordwell 1999; Galperin and Koonin 1999; Huynen et al. 1999; Morett et al. 2003; George et al. 2004; Gherardini et al. 2007; Omelchenko et al. 2010).

Several studies have suggested that the fraction of enzymatic activities in which multiple events of independent origin have taken place during evolution is substantial (Hegyi and

Gerstein 1999; Morett et al. 2003; George et al. 2004; Gherardini et al. 2007; Omelchenko et al. 2010), and some of these "missing" enzymes have been identified and characterized in some detail (Almonacid et al. 2010; Galperin and Koonin 2012). Apparently, analogous enzymes are often recruited from distinct superfamilies (Galperin et al. 1998; Omelchenko et al. 2010), with some of these alternative forms sharing the reaction catalyzed and the configuration of the catalytic residues (although these residues do not share the same fold, in these cases) (Galperin and Koonin 2012).

Despite being recognized for a long time, though erroneously referred in older literature as isozymes (or isoenzymes), isoforms, or class/type I and class/type II enzymes (e.g., Martin and Schnarrenberger 1997), functionally analogous enzymes remain poorly explored, and a comprehensive investigation of the occurrence, distribution, and implications of convergence in enzymatic activities, at least involving organisms whose genomes have been completely sequenced, has not been done so far. Fundamental questions, such as how analogous enzymes originate, why so many events of independent origin have apparently occurred during evolution, and what are the reasons for the coexistence in the same organism of distinct enzymatic forms catalyzing the same biochemical reaction, among several other questions, such as concerning the catalysis of similar reactions by different structural scaffolds (Almonacid et al. 2010), remain unanswered.

Surprisingly, numerous isofunctional enzymes are still not recognized as nonhomologous counterparts, despite substantial evidence indicating different evolutionary histories (e.g., Omelchenko et al. 2010). However, in some of these unrecognized cases, it has been demonstrated that the analogous enzymes either have an unsuspected separate evolutionary history or present (experimentally verified) distinct functional features, as we will discuss later.

In this work, we begin to investigate the biological significance of the cooccurrence of nonhomologous isofunctional enzymes in human metabolism, characterizing functional analogous enzymes identified in biochemical pathways and processes annotated in the human genome. Our hypothesis is that the coexistence of multiple enzymatic forms might not be interpreted as functional redundancy. Instead, these enzymatic forms may be implicated in distinct (and probably relevant) biological roles.

To catalog the repertoire of isofunctional enzymes cooccurring in the human metabolism (from now on referred as intragenomic analogous enzymes) a computational pipeline (AnEnPi) (Otto et al. 2008) was employed to identify putative analogous enzymes using the KEGG database (Kanehisa and Goto 2000) as the source of information. The predicted functional analogy instances were confirmed based on domain, folding and 3D structure information assigned to the enzymes implicated (see Materials and Methods for details).

Altogether, we could find evidence of convergence in 15 enzymatic activities belonging to 45 distinct processes and metabolic pathways represented in KEGG's human reference maps (table 1, and supplementary materials I and II, Supplementary Material online). The genomic coordinates of the genes encoding these predicted analogous enzymes showed that these genes are dispersed throughout the human genome, with most of the genes encoding for distinct analogous forms (as well as duplications of several alternative forms) located on separated chromosomes (fig. 2).

The Repertoire of Nonhomologous Isofunctional Enzymes in Humans

One valuable source of information on enzymatic activities and metabolic pathways is the KEGG Pathway database available at the Kyoto Encyclopedia of Genes and Genomes (KEGG) platform, which comprises a collection of manually elaborated maps representing the current knowledge about networks of molecular interaction in biological processes or biochemical pathways. Hence, from KEGG database version 73.1, we obtained 1,159,633 protein sequences encoded in 2,494 genomes, ranging all three domains of life (Archaea, Bacteria, and Eukarya), distributed in 3,825 enzymatic activities. From these enzymatic activities, 3,572 were fully annotated with the four EC digits classification, containing 1,025,885 protein sequences. On the other hand, 253 incomplete ECs were identified (defined until the first, second, or third digit of the EC classification scheme), comprising 133,748 sequences.

Different types of convergence occur at the molecular level and can be categorized into functional, mechanistic, structural, and sequence. Thus, enzymes whose chemical transformations are defined only by three digits of EC classification may have different reaction specificities (different substrates/products or cofactors), constituting mechanistic analogous, that is, unrelated enzymes which catalyze distinct chemical transformations through the same mechanism of action (Doolittle 1994; Gherardini et al. 2007). However, this sort of event is not considered in this work, which is dedicated solely to investigate functional analogy. Thus, the AnEnPi computational prediction (including all organisms and enzymatic activities in KEGG database) resulted in 2,203 enzymatic activities in which protein sequences were grouped in two or more distinct clusters, comprising 1,996 enzymatic activities with four-digit EC annotation.

Considering only the inference of convergence in enzymatic activities with four-digit EC classification annotated in the human genome, we found 150 ECs (2,288 protein sequences) identified by AnEnPi as sustaining putative events of de novo origin. After removing from our data set enzymatic activities in which protein sequences were annotated as "subunits" and "chains," as well as enzymatic activities containing clusters composed of a single human sequence or in

Table 1

Sequence and Structure Similarity Profile of Protein Sequences Comprising Each Enzymatic Activity Assigned to the Bona Fide (+) Data Set of Intragenomic Analogous Enzymes

EC	Gene	UniprotKB	PDB	Cluster	Gene	UniprotKB	PDB	Cluster	Identity (%)	Similarity (%)	Score	TM-Score	RMSD
1.3.1.20	DHDH	Q9UQ10	2O48[a]	1	AKR1C2	P52895	2HDJ	2	16.3	28.0	40.0	0.34186	6.09
	DHDH	Q9UQ10	2O48[a]	1	AKR1C1	Q04828	1J96	2	16.0	27.5	33.0	0.33888	5.89
	AKR1C2	P52895	2HDJ	2	AKR1C1	Q04828	1J96	2	97.8	98.5	1662.0	0.99487	0.40
1.15.1.1	SOD2	P04179	1LUV	1	SOD1	P00441	4XCR	2	13.6	24.8	37.5	0.24203	4.48
	SOD2	P04179	1LUV	1	SOD3	P08294	2JLP	2	3.7	6.7	21.5	0.28352	5.09
	SOD1	P00441	4XCR	2	SOD3	P08294	2JLP	2	25.1	34.4	265.5	0.70353	1.76
2.4.1.22	B4GALT2	O60909	ND	1	B4GALT1	P15291	2AH9	1	50.0	62.7	1049.5	ND	ND
	B4GALT2	O60909	ND	1	LALBA	P00709	3B0O	2	4.0	7.0	7.5	ND	ND
	B4GALT1	P15291	2AH9	1	LALBA	P00709	3B0O	2	6.8	12.0	9.0	0.23387	5.70
2.4.2.31	SIRT6	Q8N6T7	3K35	1	ART1	P52961	ND	2	15.4	23.7	30.5	ND	ND
	SIRT6	Q8N6T7	3K35	1	ART3	Q13508	ND	2	18.6	27.2	43.5	ND	ND
	SIRT6	Q8N6T7	3K35	1	ART4	Q93070	ND	2	2.1	3.2	17.5	ND	ND
	SIRT6	Q8N6T7	3K35	1	ART5	Q96L15	ND	2	4.3	5.7	15.5	ND	ND
	ART1	P52961	ND	2	ART3	Q13508	ND	2	21.7	33.0	263.5	ND	ND
	ART1	P52961	ND	2	ART4	Q93070	ND	2	29.3	42.7	377.0	ND	ND
	ART1	P52961	ND	2	ART5	Q96L15	ND	2	34.8	46.6	447.5	ND	ND
	ART3	Q13508	ND	2	ART4	Q93070	ND	2	18.8	30.5	221.0	ND	ND
	ART3	Q13508	ND	2	ART5	Q96L15	ND	2	25.6	35.1	391.5	ND	ND
	ART4	Q93070	ND	2	ART5	Q96L15	ND	2	28.7	43.9	321.0	ND	ND
2.7.1.67	PI4KA	P42356	ND	1	PI4KB	Q9UBF8	4WAE	1	10.8	17.1	527.5	ND	ND
	PI4KA	P42356	ND	1	PI4K2A	Q9BTU6	4HND	2	3.8	6.4	46.0	ND	ND
	PI4KA	P42356	ND	1	PI4K2B	Q8TCG2	4WTV	2	4.7	8.0	28.5	ND	ND
	PI4KB	Q9UBF8	4WAE	1	PI4K2A	Q9BTU6	4HND	2	9.8	16.4	45.5	0.48577	4.85
	PI4KB	Q9UBF8	4WAE	1	PI4K2B	Q8TCG2	4WTV	2	9.9	16.5	51.0	0.31144	4.58
	PI4K2A	Q9BTU6	4HND	2	PI4K2B	Q8TCG2	4WTV	2	57.7	69.5	1472.5	0.46188	1.74
2.7.4.21	PPIP5K2	O43314	3T9A	1	PPIP5K1	Q6PFW1	ND	1	56.2	64.5	4170.5	ND	ND
	PPIP5K2	O43314	3T9A	1	IP6K1	Q92551	ND	2	8.0	12.8	50.0	ND	ND
	PPIP5K2	O43314	3T9A	1	IP6K3	Q96PC2	ND	2	7.1	11.2	49.5	ND	ND
	PPIP5K2	O43314	3T9A	1	IP6K2	Q9UHH9	ND	2	7.0	11.3	47.5	ND	ND
	PPIP5K1	Q6PFW1	ND	1	IP6K1	Q92551	ND	2	6.7	10.5	96.5	ND	ND
	PPIP5K1	Q6PFW1	ND	1	IP6K3	Q96PC2	ND	2	5.8	10.2	42.5	ND	ND
	PPIP5K1	Q6PFW1	ND	1	IP6K2	Q9UHH9	ND	2	6.1	9.6	43.5	ND	ND
	IP6K1	Q92551	ND	2	IP6K3	Q96PC2	ND	2	47.6	61.4	1072.0	ND	ND
	IP6K1	Q92551	ND	2	IP6K2	Q9UHH9	ND	2	46.3	62.2	1019.0	ND	ND
	IP6K3	Q96PC2	ND	2	IP6K2	Q9UHH9	ND	2	44.7	58.7	911.0	ND	ND
3.1.1.3	AADAC	P22760	ND	3	CEL	B4DSX9	ND	3	10.3	17.7	100.5	ND	ND
	AADAC	P22760	ND	3	LIPC	P11150	ND	4	17.1	27.6	26.0	ND	ND
	AADAC	P22760	ND	3	PNLIP	P16233	1LPB	4	13.6	24.5	18.0	ND	ND
	AADAC	P22760	ND	3	PNLIPRP1	P54315	2PPL	4	11.5	20.7	25.5	ND	ND
	AADAC	P22760	ND	3	PNLIPRP3	Q17RR3	ND	4	13.5	22.3	36.0	ND	ND
	AADAC	P22760	ND	3	LIPG	Q9Y5X9	ND	4	11.7	22.4	15.0	ND	ND
	AADAC	P22760	ND	3	PNLIPRP2	P54317	2OXE	4	14.6	26.6	36.0	ND	ND
	CEL	B4DSX9	ND	3	LIPC	P11150	ND	4	10.3	15.8	41.0	ND	ND
	CEL	B4DSX9	ND	3	PNLIP	P16233	1LPB	4	11.8	18.7	34.0	ND	ND
	CEL	B4DSX9	ND	3	PNLIPRP1	P54315	2PPL	4	5.6	8.6	29.0	ND	ND
	CEL	B4DSX9	ND	3	PNLIPRP3	Q17RR3	ND	4	7.4	12.7	31.5	ND	ND
	CEL	B4DSX9	ND	3	LIPG	Q9Y5X9	ND	4	11.5	19.2	42.5	ND	ND
	CEL	B4DSX9	ND	3	PNLIPRP2	P54317	2OXE	4	4.8	8.5	21.5	ND	ND
	AADAC	P22760	ND	3	LIPF	P07098	1HLG	9	18.2	31.1	57.0	ND	ND
	CEL	B4DSX9	ND	3	LIPF	P07098	1HLG	9	13.5	22.3	32.0	ND	ND
	AADAC	P22760	ND	3	PNPLA3	Q9NST1	ND	10	3.4	5.1	36.0	ND	ND
	CEL	B4DSX9	ND	3	PNPLA3	Q9NST1	ND	10	17.8	26.5	42.5	ND	ND
	LIPC	P11150	ND	4	PNLIP	P16233	1LPB	4	28.4	42.1	503.0	ND	ND

(continued)

Table 1 Continued

EC	Gene	UniprotKB	PDB	Cluster	Gene	UniprotKB	PDB	Cluster	Identity (%)	Similarity (%)	Score	TM-Score	RMSD
	LIPC	P11150	ND	4	PNLIPRP1	P54315	2PPL	4	29.0	42.3	506.0	ND	ND
	LIPC	P11150	ND	4	PNLIPRP3	Q17RR3	ND	4	29.5	43.9	536.0	ND	ND
	LIPC	P11150	ND	4	LIPG	Q9Y5X9	ND	4	41.3	61.0	1059.5	ND	ND
	LIPC	P11150	ND	4	PNLIPRP2	P54317	2OXE	4	27.1	43.0	473.5	ND	ND
	PNLIP	P16233	1LPB	4	PNLIPRP1	P54315	2PPL	4	67.3	80.6	1750.0	0.93392	1.76
	PNLIP	P16233	1LPB	4	PNLIPRP3	Q17RR3	ND	4	47.3	63.7	1113.5	ND	ND
	PNLIP	P16233	1LPB	4	LIPG	Q9Y5X9	ND	4	30.5	42.2	556.5	ND	ND
	PNLIP	P16233	1LPB	4	PNLIPRP2	P54317	2OXE	4	64.0	79.5	1676.0	0.94537	1.37
	PNLIPRP1	P54315	2PPL	4	PNLIPRP3	Q17RR3	ND	4	48.4	64.3	1158.0	ND	ND
	PNLIPRP1	P54315	2PPL	4	LIPG	Q9Y5X9	ND	4	28.5	42.5	543.5	ND	ND
	PNLIPRP1	P54315	2PPL	4	PNLIPRP2	P54317	2OXE	4	62.7	77.0	1655.0	0.92898	1.81
	PNLIPRP3	Q17RR3	ND	4	LIPG	Q9Y5X9	ND	4	29.3	44.2	519.0	ND	ND
	PNLIPRP3	Q17RR3	ND	4	PNLIPRP2	P54317	2OXE	4	47.8	62.2	1156.5	ND	ND
	LIPG	Q9Y5X9	ND	4	PNLIPRP2	P54317	2OXE	4	28.6	44.9	536.5	ND	ND
	LIPC	P11150	ND	4	LIPF	P07098	1HLG	9	15.9	28.2	41.0	ND	ND
	PNLIP	P16233	1LPB	4	LIPF	P07098	1HLG	9	17.2	26.2	65.5	0.39307	4.45
	PNLIPRP1	P54315	2PPL	4	LIPF	P07098	1HLG	9	16.8	29.2	37.0	0.38582	5.08
	PNLIPRP3	Q17RR3	ND	4	LIPF	P07098	1HLG	9	16.5	25.8	41.5	ND	ND
	LIPG	Q9Y5X9	ND	4	LIPF	P07098	1HLG	9	12.3	22.0	26.0	ND	ND
	PNLIPRP2	P54317	2OXE	4	LIPF	P07098	1HLG	9	6.1	9.8	26.0	0.40059	5.14
	LIPC	P11150	ND	4	PNPLA3	Q9NST1	ND	10	6.9	11.9	48.5	ND	ND
	PNLIP	P16233	1LPB	4	PNPLA3	Q9NST1	ND	10	11.6	18.8	31.0	ND	ND
	PNLIPRP1	P54315	2PPL	4	PNPLA3	Q9NST1	ND	10	14.3	21.2	54.0	ND	ND
	PNLIPRP3	Q17RR3	ND	4	PNPLA3	Q9NST1	ND	10	9.2	16.5	34.0	ND	ND
	LIPG	Q9Y5X9	ND	4	PNPLA3	Q9NST1	ND	10	7.9	12.3	30.0	ND	ND
	PNLIPRP2	P54317	2OXE	4	PNPLA3	Q9NST1	ND	10	13.1	20.8	38.5	ND	ND
	LIPF	P07098	1HLG	9	PNPLA3	Q9NST1	ND	10	2.5	3.5	7.0	ND	ND
3.1.1.29	PTRH2	Q9Y3E5	1Q7S	1	PTRH1	Q86Y79	ND	2	17.6	27.5	28.0	ND	ND
	PTRH2	Q9Y3E5	1Q7S	1	ICT1	Q14197	ND	3	10.2	14.9	18.0	ND	ND
	PTRH1	Q86Y79	ND	2	ICT1	Q14197	ND	3	13.3	21.8	15.5	ND	ND
3.1.2.2	ACOT2	P49753	3HLK	1	BAAT	Q14032	ND	1	38.0	51.0	873.5	ND	ND
	ACOT2	P49753	3HLK	1	ACOT1	Q86TX2	ND	1	86.1	86.5	2217.0	ND	ND
	ACOT2	P49753	3HLK	1	ACOT4	Q8N9L9	3K2I	1	61.1	70.6	1601.0	0.95168	1.20
	BAAT	Q14032	ND	1	ACOT1	Q86TX2	ND	1	42.9	56.9	868.5	ND	ND
	BAAT	Q14032	ND	1	ACOT4	Q8N9L9	3K2I	1	43.1	56.7	841.0	ND	ND
	ACOT1	Q86TX2	ND	1	ACOT4	Q8N9L9	3K2I	1	70.3	81.0	1603.0	ND	ND
	ACOT2	P49753	3HLK	1	ACOT7	O00154	2QQ2	2	2.8	4.6	27.5	0.19192	4.25
	BAAT	Q14032	ND	1	ACOT7	O00154	2QQ2	2	0.4	0.8	9.0	ND	ND
	ACOT1	Q86TX2	ND	1	ACOT7	O00154	2QQ2	2	13.4	21.5	18.5	ND	ND
	ACOT4	Q8N9L9	3K2I	1	ACOT7	O00154	2QQ2	2	2.0	2.5	13.5	0.23303	5.39
3.1.3.2	ACP5	P13686	1WAR	1	ACP2	P11117	ND	2	4.2	7.9	23.0	ND	ND
	ACP5	P13686	1WAR	1	ACPP	P15309	1CVI	2	16.6	28.3	22.5	0.38800	5.44
	ACP5	P13686	1WAR	1	ACPT	Q9BZG2	ND	2	17.4	24.4	43.0	ND	ND
	ACP5	P13686	1WAR	1	ACP6	Q9NPH0	4JOB	2	15.5	24.3	19.5	0.36502	5.59
	ACP5	P13686	1WAR	1	ACP1	P24666	5PNT	5	10.1	17.0	24.5	0.34702	5.16
	ACP2	P11117	ND	2	ACPP	P15309	1CVI	2	43.6	58.4	976.5	ND	ND
	ACP2	P11117	ND	2	ACPT	Q9BZG2	ND	2	43.0	57.2	842.5	ND	ND
	ACP2	P11117	ND	2	ACP6	Q9NPH0	4JOB	2	21.3	33.9	269.5	ND	ND
	ACPP	P15309	1CVI	2	ACPT	Q9BZG2	ND	2	36.8	50.2	770.0	ND	ND
	ACPP	P15309	1CVI	2	ACP6	Q9NPH0	4JOB	2	26.1	41.5	319.5	0.80111	2.77
	ACPT	Q9BZG2	ND	2	ACP6	Q9NPH0	4JOB	2	24.4	35.3	289.5	ND	ND
	ACP2	P11117	ND	2	ACP1	P24666	5PNT	5	8.5	13.9	28.5	ND	ND
	ACPP	P15309	1CVI	2	ACP1	P24666	5PNT	5	10.1	18.9	19.0	0.26430	5.89
	ACPT	Q9BZG2	ND	2	ACP1	P24666	5PNT	5	6.0	12.9	13.0	ND	ND
	ACP6	Q9NPH0	4JOB	2	ACP1	P24666	5PNT	5	2.2	4.1	16.5	0.24707	4.77

(continued)

Table 1 Continued

EC	Gene	UniprotKB	PDB	Cluster	Gene	UniprotKB	PDB	Cluster	Identity (%)	Similarity (%)	Score	TM-Score	RMSD
3.1.3.5	NT5C1B	Q96P26	ND	2	NT5C1A	Q9BXI3	ND	2	35.4	43.0	1145.0	ND	ND
	NT5C1B	Q96P26	ND	2	NT5E	P21589	4H2G	3	8.5	13.3	26.0	ND	ND
	NT5C1A	Q9BXI3	ND	2	NT5E	P21589	4H2G	3	10.8	20.1	32.5	ND	ND
	NT5C1B	Q96P26	ND	2	NT5C	Q8TCD5	4L57	5	7.3	13.2	29.5	ND	ND
	NT5C1B	Q96P26	ND	2	NT5M	Q9NPB1	4MUM	5	6.0	9.3	48.0	ND	ND
	NT5C1A	Q9BXI3	ND	2	NT5C	Q8TCD5	4L57	5	9.7	15.1	29.0	ND	ND
	NT5C1A	Q9BXI3	ND	2	NT5M	Q9NPB1	4MUM	5	8.5	14.0	26.5	ND	ND
	NT5C1B	Q96P26	ND	2	NT5C3A	Q9H0P0	2CN1	7	8.6	16.6	35.5	ND	ND
	NT5C1A	Q9BXI3	ND	2	NT5C3A	Q9H0P0	2CN1	7	9.1	17.1	23.5	ND	ND
	NT5C1B	Q96P26	ND	2	NT5C2	P49902	2XCW	9	8.6	13.9	42.5	ND	ND
	NT5C1A	Q9BXI3	ND	2	NT5C2	P49902	2XCW	9	8.2	14.5	20.0	ND	ND
	NT5E	P21589	4H2G	3	NT5C	Q8TCD5	4L57	5	7.7	13.4	7.0	0.26104	5.86
	NT5E	P21589	4H2G	3	NT5M	Q9NPB1	4MUM	5	2.0	3.2	12.0	0.26951	5.69
	NT5E	P21589	4H2G	3	NT5C3A	Q9H0P0	2CN1	7	8.9	15.9	30.5	0.26314	5.72
	NT5E	P21589	4H2G	3	NT5C2	P49902	2XCW	9	15.4	26.8	36.0	0.27774	7.25
	NT5C	Q8TCD5	4L57	5	NT5M	Q9NPB1	4MUM	5	51.3	64.7	660.0	0.96844	0.70
	NT5C	Q8TCD5	4L57	5	NT5C3A	Q9H0P0	2CN1	7	6.7	9.9	27.5	0.50279	4.85
	NT5M	Q9NPB1	4MUM	5	NT5C3A	Q9H0P0	2CN1	7	14.1	26.9	40.0	0.50929	4.86
	NT5C	Q8TCD5	4L57	5	NT5C2	P49902	2XCW	9	7.5	11.8	26.0	0.44003	3.79
	NT5M	Q9NPB1	4MUM	5	NT5C2	P49902	2XCW	9	4.6	9.1	23.5	0.44594	3.87
	NT5C3A	Q9H0P0	2CN1	7	NT5C2	P49902	2XCW	9	8.5	14.7	46.0	0.45164	4.71
3.1.4.12	SMPD2	O60906	ND	1	SMPD3	Q9NY59	ND	1	10.4	15.2	87.5	ND	ND
	SMPD2	O60906	ND	1	SMPD1	P17405	5I81	2	5.5	10.0	35.5	ND	ND
	SMPD3	Q9NY59	ND	1	SMPD1	P17405	5I81	2	2.6	3.6	51.0	ND	ND
	SMPD2	O60906	ND	1	SMPD4	Q9NXE4	ND	3	8.2	13.6	62.0	ND	ND
	SMPD3	Q9NY59	ND	1	SMPD4	Q9NXE4	ND	3	12.7	20.2	51.5	ND	ND
	SMPD2	O60906	ND	1	ENPP7	Q6UWV6	5UDY	4	7.0	12.8	27.5	ND	ND
	SMPD3	Q9NY59	ND	1	ENPP7	Q6UWV6	5UDY	4	13.3	19.2	29.5	ND	ND
	SMPD1	P17405	5I81	2	SMPD4	Q9NXE4	ND	3	6.3	10.7	39.0	ND	ND
	SMPD1	P17405	5I81	2	ENPP7	Q6UWV6	5UDY	4	16.1	25.0	28.5	0.28698	6.61
	SMPD4	Q9NXE4	ND	3	ENPP7	Q6UWV6	5UDY	4	5.6	8.7	49.5	ND	ND
4.2.99.18	NTHL1	P78549	ND	1	OGG1	O15527	1KO9	1	19.0	28.4	85.5	ND	ND
	NTHL1	P78549	ND	1	NEIL2	Q969S2	1VZP	1	14.6	25.7	29.0	ND	ND
	NTHL1	P78549	ND	1	NEIL1	Q96FI4	1TDH	1	6.7	9.8	40.0	ND	ND
	OGG1	O15527	1KO9	1	NEIL2	Q969S2	1VZP	1	2.9	4.3	24.0	0.27899	4.84
	OGG1	O15527	1KO9	1	NEIL1	Q96FI4	1TDH	1	14.3	20.2	42.5	0.28166	6.65
	NTHL1	P78549	ND	1	APEX1	P27695	2O3H	1	15.6	23.0	41.0	ND	ND
	NTHL1	P78549	ND	1	APEX2	Q9UBZ4	ND	1	4.7	6.6	14.5	ND	ND
	OGG1	O15527	1KO9	1	APEX1	P27695	2O3H	1	16.5	26.7	23.5	0.27998	6.74
	OGG1	O15527	1KO9	1	APEX2	Q9UBZ4	ND	1	9.4	14.6	42.5	ND	ND
	NTHL1	P78549	ND	1	APLF	Q8IW19	2KUO	6	3.7	6.1	24.5	ND	ND
	OGG1	O15527	1KO9	1	APLF	Q8IW19	2KUO	6	6.0	8.7	20.5	0.14656	6.94
	NEIL2	Q969S2	1VZP	1	NEIL1	Q96FI4	1TDH	1	18.8	24.9	143.5	0.50213	2.44
	NEIL2	Q969S2	1VZP	1	APEX1	P27695	2O3H	1	13.2	22.2	19.5	0.28098	4.80
	NEIL2	Q969S2	1VZP	1	APEX2	Q9UBZ4	ND	1	6.6	10.5	42.0	ND	ND
	NEIL1	Q96FI4	1TDH	1	APEX1	P27695	2O3H	1	5.2	7.3	44.0	0.25323	6.80
	NEIL1	Q96FI4	1TDH	1	APEX2	Q9UBZ4	ND	1	14.6	20.8	41.0	ND	ND
	NEIL2	Q969S2	1VZP	1	APLF	Q8IW19	2KUO	6	3.9	7.9	24.5	0.09438	5.32
	NEIL1	Q96FI4	1TDH	1	APLF	Q8IW19	2KUO	6	10.7	18.3	46.5	0.15631	6.56
	APEX1	P27695	2O3H	1	APEX2	Q9UBZ4	ND	1	14.9	22.0	264.0	ND	ND
	APEX1	P27695	2O3H	1	APLF	Q8IW19	2KUO	6	6.7	12.8	47.5	0.16022	6.62
	APEX2	Q9UBZ4	ND	1	APLF	Q8IW19	2KUO	6	12.3	20.5	44.5	ND	ND
5.3.99.2	PTGDS	P41222	2WWP	1	HPGDS	O60760	1IYI	2	13.7	21.4	9.5	0.28983	5.76
5.3.99.3	PTGES2	Q9H7Z7	ND	1	PTGES	O14684	4AL0	2	0.6	0.6	18.0	ND	ND

[a]The 3D model of the human dehydrogenase (UniProt Q9UQ10) was obtained using the crystal structure of a *Macaca fascicularis* dehydrogenase (PDB 2O48) by comparative modeling.

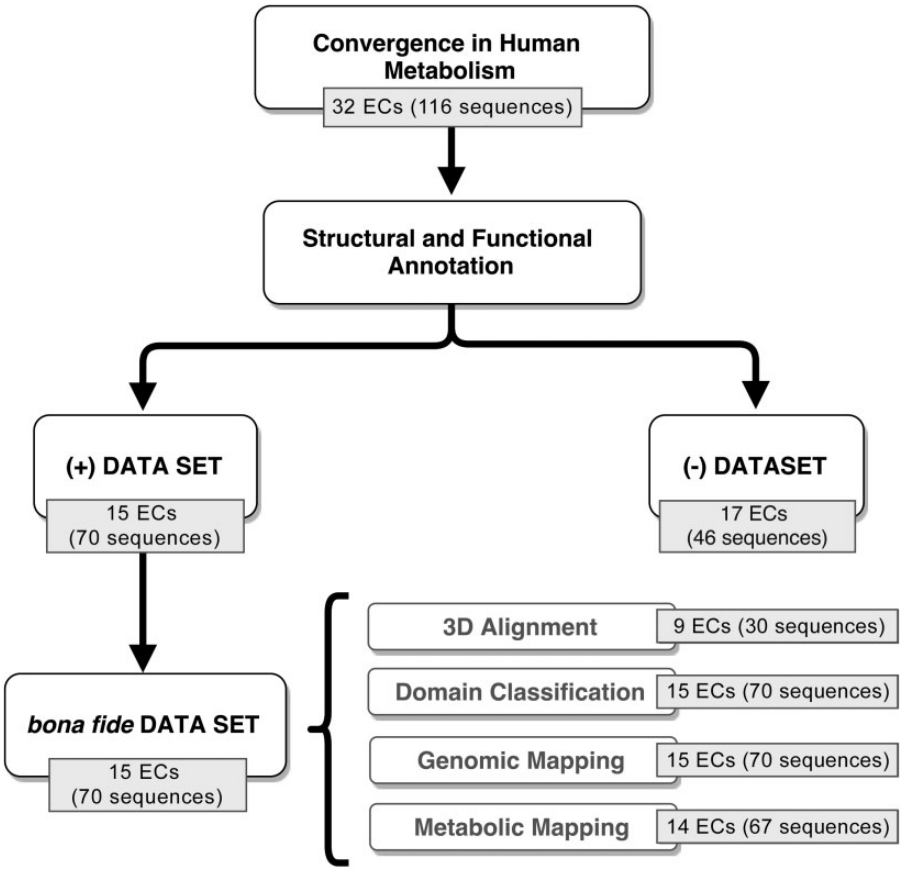


Fig. 1.—Outline of the procedure used for the identification of intragenomic analogy in human metabolism (see text for details).

which the protein sequences were grouped in one single cluster (see Materials and Methods for details), we obtain 116 protein sequences comprising 32 distinct enzymatic activities in human metabolism.

A flowchart representing our downstream data analyses is shown in figure 1. Overall, 116 protein sequences were initially predicted as pairs or groups of alternative forms in 32 enzymatic activities of the human metabolism. From these, 70 protein sequences, comprising 15 ECs, were assigned to the bona fide (+) data set, in which all enzymatic activities are composed of putative alternative enzymatic forms belonging to at least two distinct superfamilies. The remaining 46 sequences (17 ECs), assigned to the (−) data set, were rejected from our analysis.

Hydrolase class (36 protein sequences in 6 ECs) was the most frequent class in our bona fide (+) data set, followed by Transferases (17 protein sequences in four ECs), Oxidoreductases (six protein sequences in two ECs), Lyases (seven protein sequences in one EC), and Isomerases (four protein sequences in two ECs). No evidence of convergence was found in Ligase class. On the other hand, these 15 enzymatic activities are mapped in 45 biochemical pathways or processes of several major metabolic classes: Aging, Cancers, Carbohydrate metabolism, Cellular community—eukaryotes, Development, Digestive system, Endocrine system, Glycan biosynthesis and metabolism, Immune diseases, Lipid metabolism, Metabolism of cofactors and vitamins, Metabolism of other amino acids, Neurodegenerative diseases, Nucleotide metabolism, Replication and repair, Signal transduction, Transport and catabolism, Xenobiotics biodegradation and metabolism (supplementary material I, Supplementary Material online).

It is worth noticing that 12 of these 15 enzymatic activities in the bona fide (+) data set (∼73%) were previously reported as presenting evidence of analogy (Capriles et al. 2010; Omelchenko et al. 2010): 1.3.1.20 (Trans-1,2-dihydrobenzene-1,2-diol dehydrogenase), 1.15.1.1 (Superoxide dismutase), 2.7.4.21 (Inositol-hexakisphosphate kinase), 3.1.1.29 (Aminoacyl-tRNA hydrolase), 3.1.2.2 (Palmitoyl-CoA hydrolase), 3.1.3.2 (Acid phosphatase), 3.1.3.5 (5′-nucleotidase), 3.1.4.12 (Sphingomyelin phosphodiesterase), 4.2.99.18 (DNA-(apurinic or apyrimidinic site) lyase), 5.3.99.2 (Prostaglandin-D synthase), 5.3.99.3 (Prostaglandin-E synthase), and 3.1.1.3 (Triacylglycerol lipase).

The SUPERFAMILY database (Wilson et al. 2009) consists of a collection of hidden Markov models, representing structural protein domains according to SCOP superfamily classification. Consequently, a superfamily groups together domains which

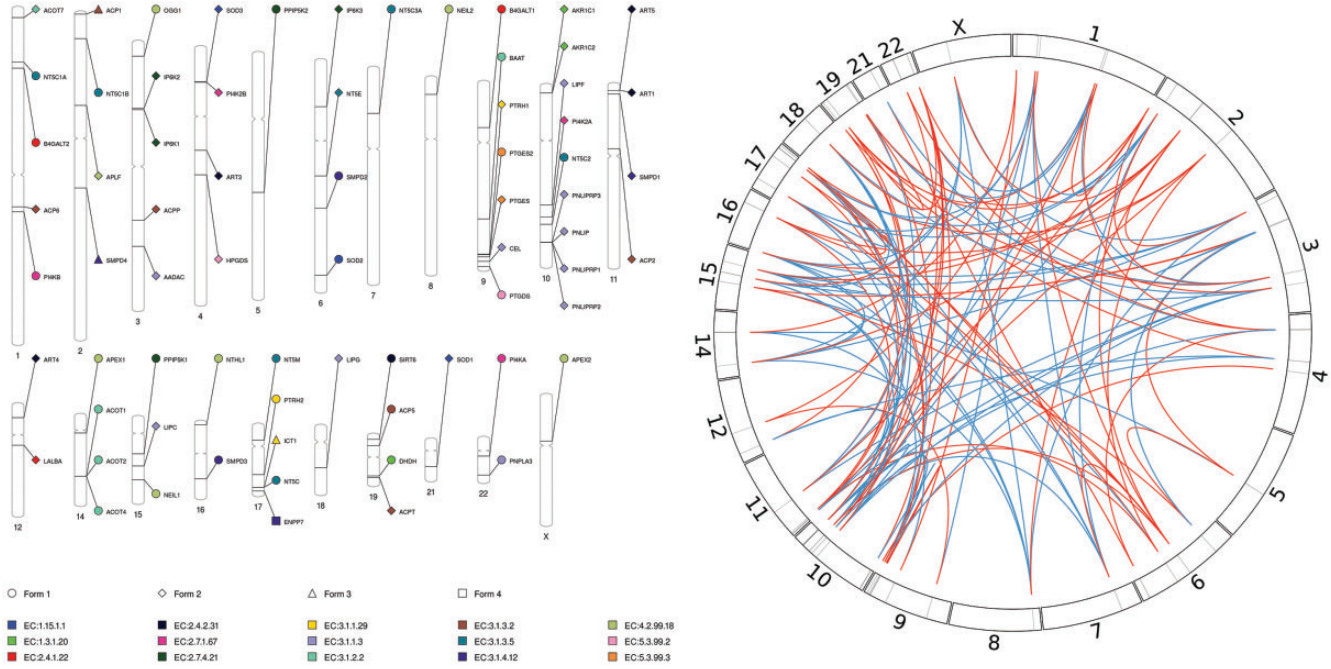


Fig. 2.—(Left) Diagram depicting the localization of genes encoding intragenomic analogous enzymes across the human chromosome. Enzymatic activities in which evidence of intragenomic analogy was found are represented by distinct colors. Genes encoding distinct enzymatic forms are represented by different symbols. (Right) Circular diagram representing the distances between genes encoding alternative forms (distinct AnEnPi cluster of the same EC) as red lines, and genes encoding homologous enzymatic forms (belonging to the same AnEnPi cluster-EC group) as blue lines. Human chromosomes are depicted as contiguous segments in a circle, in which vertical black bars along the extension of these segments (chromosomes) represent the location of the 70 genes encoding intragenomic analogous enzymes comprising our bona fide (+) data set. Short lines (red and blue) represent neighbor genes in a chromosome.

have an evolutionary relationship. Hence, considering the SUPERFAMILY classification, we identified 39 different superfamilies (38 distinct folds) among the putative analogous enzymes in the bona fide (+) data set. The most frequent superfamilies are: alpha/beta-Hydrolases (13), Phosphoglycerate mutase-like (6), Lipase/lipooxygenase domain (PLAT/LH2 domain) (6), ADP-ribosylation (5), DNase I-like (4), HAD-like (4), Metallo-dependent phosphatases (3), SAICAR synthase-like (3), NAD(P)-linked oxidoreductase (2), Protein kinase-like (PK-like) (2), Cu,Zn superoxide dismutase-like (2), DNA-glycosylase (2), Glutathione synthetase ATP-binding domain-like (2), GST C-terminal domain-like (2), Nucleotide-diphospho-sugar transferases (2), S13-like H2TH domain (2), and Thioredoxin-like (2), followed by 22 different superfamilies represented once. On the other hand, we identified 51 distinct Pfam domains/families in those 70 enzymes. Of these, 29 enzymes are multidomain and 41 are composed of (or annotated as) a single domain. Three domains are shared among some enzymatic activities: His_Phos_2 (ECs 2.7.4.21 and 3.1.3.2), Metallophos (ECs 3.1.3.2, 3.1.3.5 and 3.1.4.12), and Exo_endo_phos (ECs 3.1.4.12 and 4.2.99.18). With two exceptions, enzymatic forms assigned to separate AnEnPi clusters have correspondingly different domain composition, indicating that inside a particular cluster-EC group, sequences might share a common origin. However, alternative forms of the enzymatic activity 2.7.1.67 display the same domain composition (PI3_PI4_kinase), although they are members of unrelated superfamilies (ARM repeat, Protein kinase-like (PK-like), and ADP-ribosylation). Enzymatic activity 4.2.99.18, on the other hand, exhibits a much more complex pattern of domain and superfamily composition (supplementary material I, Supplementary Material online).

To measure the similarity among these 70 sequences in the bona fide (+) data set, we performed a global rigorous pairwise sequence alignment (table 1). The highest score, similarity, and identity values were observed between enzymatic forms belonging to the same AnEnPi cluster, as expected, since enzymes that share the same enzymatic activity, grouped in the same cluster, are presumably homologous. We obtained similar results when 3D structures of these protein sequences were compared, employing the TM-score (Zhang and Skolnick 2004) and RMSD (root-mean-square deviation of atomic positions) measurements to estimate the similarity between them. We applied the following thresholds to distinguish related and unrelated structures: TM-score-< 0.2, indicating a probable distinct evolutionary origin, and TM-score > 0.5, mostly corresponding to the same fold in SCOP (Murzin et al. 1995) or CATH (Sillitoe et al. 2015). Most of the alternative forms obtained TM-scores < 0.5

when their structures were aligned (table 1 and supplementary material II, Supplementary Material online). Therefore, comparisons between sequences belonging to the same AnEnPi cluster-EC group resulted in RMSD values tending to zero and TM-scores close to 1, indicating a possible common evolutionary origin. When sequences belonging to distinct clusters of the same EC were compared, the opposite trend was observed, as expected (table 1). The intermediate TM-scores observed between the products of the genes NT5C3A and the alternative forms encoded by genes NT5C (0.50279) and NT5M (0.50929), as well as between the products of the genes CEL and LIPF (0.47684), can be attributed to the folds HAD-like and alpha/beta-Hydrolases shared between them, respectively.

In summary, we could assess the inference of convergence in all those 15 enzymatic activities based on superfamily and domain information, and based on structural alignments between the predicted alternative forms in 9 out of 15 of those enzymatic activities as well (ECs 1.3.1.20, 1.15.1.1, 2.4.1.22, 2.7.1.67, 3.1.2.2, 3.1.3.2, 3.1.3.5, 4.2.99.18, and 5.3.99.2). As shown in figure 2*A*, except for genes PTGES and PTGES2 encoding enzymes of the enzymatic activity 5.3.99.3, on chromosome 9 the genes encoding intragenomic analogous enzymes appear to be randomly distributed, dispersed throughout the entire human genome and recognized in 21 of the 24 nuclear chromosomes (20 autosomes and one sex chromosome). For genes encoding alternative forms as well as genes encoding homologous enzymatic forms, we mapped the chromosomal locations and then plotted in a circular diagram, as shown in figure 2*B*. Likewise, the distances between genes encoding intragenomic analogous and between homologous enzymes, exhibit a similar fuzzy pattern of occurrence in the human genome.

Nucleotidades, Dehydrogenases, Synthases, Dismutases, Kinases, and Lipases

The literature indicates the existence of seven human 5′-nucleotidases (EC 3.1.3.5), hydrolases involved in the biosynthesis of nucleosides and inorganic phosphate from noncyclic nucleoside monophosphates, encoded by the genes NT5E, NT5C1A, NT5C1B, NT5C, NT5C3A, NT5C2, and NT5M: one soluble enzyme associated with the cell membrane (NT5E), and six enzymes with an intracellular location, either cytosolic (NT5C1A, NT5C1B, NT5C, NT5C3A, and NT5C2) or mitochondrial (NT5M) (Zukowska et al. 2015). All these genes are distributed in several distinct chromosomes (1, 2, 6, 7, 10, and 17) (fig. 2), and the enzymes encoded by them were assigned to five AnEnPi clusters: 1) NT5E, 2) NT5C1A and NT5C1B, 3) NT5C and NT5M, 4) NT5C3A, and 5) NT5C2 (supplementary material I, Supplementary Material online). The enzyme encoded by the gene NT5E belongs to the 5′-nucleotidase, C-terminal domain, and Metallo-dependent phosphatases superfamily, whereas the one encoded by the gene NT5C, as well as most of the remaining human 5′-nucleotidases (NT5C3A, NT5C2, NT5M), are members of the HAD-like superfamily. The enzymes encoded by the genes NT5C1A and NT5C1B do not have any superfamily annotation or any available 3D structure but were grouped together in a separate AnEnPi cluster showing considerable sequence similarity, indicating a possible common origin (table 1). The membrane-bound enzyme, NT5E, clearly distinguishes from the remaining enzymatic forms in all measures, as it was allocated in a separate AnEnPi cluster, showing remarkably low sequence and structural similarity when compared with all other 5′-nucleotidases, an entirely different superfamily/fold classification (as mentioned before), and a distinct domain composition/architecture (table 1 and supplementary material I, Supplementary Material online). On the other hand, the cytosolic, HAD-like superfamily enzymes, encoded by genes NT5C, NT5C3A, and NT5C2, as well as the mitochondrial enzyme, encoded by the NT5M gene, were assigned to three separate AnEnPi clusters; NT5C and NT5M enzymes, residing in the same AnEnPi cluster, show high sequence and structural similarity between them, as well as the same domain composition/architecture, whereas the opposite trend is observed when HAD-like superfamily enzymes representatives of distinct AnEnPi clusters are compared (both NT5C or NT5M against NT5C3A or NT5C2, and NT5C3A against NT5C2): very low sequence and structural similarity, and unrelated domain composition/architecture (table 1 and supplementary material I, Supplementary Material online). Interestingly, Crisp et al. (2015) showed considerable evidence that genes NT5C and NT5M had been horizontally acquired in the human lineage (possibly from bacterial genomes), therefore contributing to biochemical diversification of 5′-nucleotidases during animal evolution. Besides the diversity of subcellular localization, possible evolutionary origin, amino acid sequence, fold, and domain composition/architecture, these enzymes use 5′-nucleotides from various sources, displaying significant differences in the range of substrates (partially overlapping), as well as in substrate specificity (Zimmermann 1992). Hence, it is reasonable to think of the possibility of these enzymes fulfill different biological roles while regulating diverse physiological processes.

The oxidoreductases Trans-1,2-Dihydrobenzene-1,2-Diol Dehydrogenase (EC 1.3.1.20) comprises the enzymes encoded by the genes DHDH, AKR1C1, and AKR1C2. In our analyses, AnEnPi assigned the product of the gene DHDH to a separate cluster, whereas all remaining enzymes (encoded by AKR1C1 and AKR1C2 genes) were grouped in a different cluster. The predicted alternative forms could be distinguished based on domain composition/architecture, superfamily classification, as well as 3D structure (table 1, and supplementary materials I and II, Supplementary Material online). It is worth noticing that DHDH gene is located on chromosome 19, whereas the remaining genes are all neighbors,

colocated in chromosome 10, with their products presenting almost identical amino acid sequences (97.8% identity over the entire sequences), therefore reinforcing the evidence of common origin (possibly recent duplication) for the genes AKR1C1 and AKR1C2 (fig. 2). The low sequence and structural similarity between DHDH enzyme and members of the aldo-keto reductase family (e.g., AKR1C1 and AKR1C2 enzymes) has already been reported, as well as differences in use of substrates (Arimitsu et al. 1999; Carbone et al. 2008). DHDH enzyme acts on (–)-[1R,2R]-dihydrodiols, while aldo-keto reductases oxidize (+)-[1S,2S]-dihydrodiols (Carbone et al. 2008). Also, aldo-keto reductase members use synthetic steroids as substrate (Penning et al. 2015). We are aware that homologous enzymes can also present distinct substrate specificities, but in this case, the established substrate difference clearly correlates with the assumed separate evolutionary origin, even in the absence of further information that could indicate other possible implication(s) in distinct biological roles.

Representatives of the enzymatic class isomerase, prostaglandin D2 synthase and hematopoietic prostaglandin D synthase (encoded by the genes PTGDS, located on chromosome 9, and HPGDS, located on chromosome 4, respectively) (EC 5.3.99.2), both regulate the synthesis of prostaglandin D2, acting in signaling and inflammatory processes (Trimarco et al. 2014; Urade and Eguchi 2002; Lim et al. 2013). Our computational pipeline AnEnPi assigned PTGDS and HPGDS enzymes to separate clusters, and subsequent analyses revealed that these enzymes are also unrelated based on domain composition, superfamily classification, as well as amino acid sequence and 3D structure (table 1, and supplementary materials I and II, Supplementary Material online), corroborating earlier evidence of functional convergence in this enzymatic activity (Urade and Eguchi 2002; Lim et al. 2013). Accordingly, an exam of the literature reveals numerous features that could clearly distinguish distinct roles for these enzymes, such as 1) the presence of signal peptide and N-glycosylation sites only in PTGDS enzyme (Urade and Eguchi 2002); 2) distinct tissue location, inhibitors, and activators, which could be related to different mechanisms of action (Urade and Eguchi 2002); 3) PTGDS enzyme is secreted, and is preferentially expressed in the brain, and is also involved in the regulation of sleep, adipogenesis, allergic and inflammatory response (Bridges et al. 2012; Marín-Méndez et al. 2012; Trimarco et al. 2014); 4) HPGDS enzyme is present in cells of the immune system (Tanaka et al. 2000).

Another major enzymatic activity in all living beings is the (oxidoreductase) superoxide dismutase (SOD) (EC 1.15.1.1); SOD enzymes catalyze the conversion of superoxide radicals (O_2^-) into hydrogen peroxide (H_2O_2) or molecular oxygen (O_2), protecting cells, tissues, and organs from oxidative stress. Humans and all other mammals express three forms of SOD: SOD1, cytoplasmatic copper/zinc enzyme (encoded by SOD1 gene on chromosome 21); SOD2, mitochondrial manganese-dependent enzyme (encoded by SOD2 gene on chromosome 6); and SOD3, extracellular copper/zinc enzyme (encoded by SOD3 gene on chromosome 4) (Landis and Tower 2005). In our computational prediction, SOD1 and SOD3 were grouped in the same AnEnPi cluster while SOD2 were assigned to a separated cluster, indicating one possible event of de novo origin. In subsequent analyses of domain composition, superfamily classification, amino acid sequence and 3D structure (table 1, and supplementary materials I and II, Supplementary Material online), we confirmed that these enzymes are indeed unrelated, corroborating previous evidence of functional convergence in SOD enzymatic activity (Omelchenko et al. 2010). In a recent study, Garcia et al. (2017) demonstrated that manganese-dependent enzymes with superoxide dismutase activity, SodA and SodM, not only coexist in the human pathogen *Staphylococcus aureus* but also clearly display distinct biological roles, in which solely one of the alternative forms, SodM, can promote resistance to antibiotics and host immunity. The authors showed that SodA is strictly manganese-dependent and relevant for combatting oxidative stress as well as for disease development when manganese is abundant, whereas SodM is truly cambialistic, essential under manganese-deplete conditions, maintaining equal enzymatic activity in the presence of manganese or iron (Garcia et al. 2017). Even though this phenomenon has only been demonstrated in bacteria so far, it opens the opportunity to explore it in other prokaryotic or eukaryotic species.

Members of the class transferase, enzymes with 1-phosphatidylinositol 4-kinase activity (PI4Ks) (EC 2.7.1.67) participate in inositol phosphate metabolism and phosphatidylinositol signaling system, catalyzing the phosphorylation of phosphatidylinositol. The product of this reaction is phosphatidylinositol 4-phosphate, a primary precursor in the synthesis of phosphatidylinositolpolyphosphates, molecules involved in many biologic processes, such as signal transduction, membrane trafficking, and cytoskeletal reorganization (Barylko et al. 2001). The mammalian PI4Ks have been classified into two types, II and III, based on physicochemical characteristics, and the literature highlights the existence of different domain organizations between PI4Ks of type II (genes PI4K2A and PI4K2B) and PI4Ks of type III (genes PI4KA and PI4KB), with PI4KA and PI4KB being more similar to each other, and PI4KA bearing a characteristic binding domain (Boura and Nencka 2015; Heilmeyer et al. 2003). Hence, the division of human PI4Ks in two separate AnEnPi clusters of putative isofunctional nonhomologous forms—one of these clusters formed by the products of the genes PI4KA and PI4KB, and the other one comprising PI4K2A and PI4K2B gene products, corresponding to the mammalian PI4Ks of type III and II, respectively, as well as their assignment to distinct superfamily classes (except for the enzyme encoded by the gene PI4K2B which has no superfamily classification) and unrelated 3D structures (supplementary materials I and II,

Supplementary Material online), reinforces similar results obtained in previous studies concerning this enzymatic activity (reviewed by Boura and Nencka 2015).

Overall, the all-against-all pairwise sequence comparison among protein sequences of each enzymatic activity of the bona fide (+) data set corroborated the AnEnPi computational predictions. However, we found at least two cases in which AnEnPi's clustering method may have "produced" more human enzymatic forms than expected: EC 3.1.3.5 (5′-nucleotidase), with five clusters, and EC 3.1.1.3 (Triacylglycerol Lipase), with four clusters. In the EC 3.1.3.5, enzymes encoded by the genes NT5C, NT5M, NT5C3A, and NT5C2 (cytoplasmic forms) share the same superfamily class (HAD-like), whereas the enzyme encoded by the gene NT5E (membrane form) is simultaneously classified in two superfamiles: 5′-nucleotidase (syn. UDP-sugar hydrolase), C-terminal domain and Metallo-dependent phosphatases. Similarly, the products of the genes AADAC, CEL, LIPC, PNLIP, PNLIPRP1, PNLIPRP3, LIPG, and LIPF, comprising the EC 3.1.1.3, are all assigned to the alpha/beta-Hydrolases superfamily, whereas the product of the PNPLA3 gene belongs to a distinct superfamily (FabD/lysophospholipase-like). Another piece of evidence supporting this assumption is that the genes PNLIPRP1, PNLIPRP2, PNLIPRP3, PNLIP, and LIPF are all neighbors, located on human chromosome 10, and their corresponding enzymes share considerably higher sequence similarity among them, than with PNPLA3, possibly representing duplication events (fig. 2 and table 1).

As we expected, all nonhomologous isofunctional enzymes we could characterize are assigned to distinct KEGG orthologous groups (KOs; http://www.genome.jp/kegg/ko.html), corroborating the distinct evolutionary origin for the predicted alternative forms (supplementary material I, Supplementary Material online). Only six KOs are shared between two or more sequences in our bona fide (+) data set: K01081, grouping six out of seven 5′-nucleotidases (NT5C1B, NT5C1A, NT5C, NT5M, NT5C3A, and NT5C2); K01046, including two out of ten triacylglycerol lipases (LIPC, and LIPG); K01068, gathering three out of five palmitoyl-CoA hydrolases (ACOT2, ACOT1, and ACOT4); K07756 and K13024 containing all five inositol-hexakisphosphate kinases (IP6K1, IP6K3, IP6K2, and PPIP5K2, PPIP5K1, respectively); and K13711, grouping two out of four 1-phosphatidylinositol 4-kinases (PI4K2A and PI4K2B).

After an extensive literature search, we were unable to find further information that could indicate (or not) a distinct evolutionary origin for the alternative forms of the remaining nine enzymatic activities of our bona fide (+) data set, neither their possible implication in distinct biological roles: EC 2.4.1.22 (Lactose synthase), EC 2.4.2.31 (NAD+—protein-arginine ADP-ribosyltransferase), EC 2.7.4.21 (Inositol-hexakisphosphate kinase), EC 3.1.1.29 (Aminoacyl-tRNA hydrolase), EC 3.1.1.3 (Triacylglycerol lipase), EC 3.1.2.2 (Palmitoyl-CoA hydrolase), EC 3.1.3.2 (Acid phosphatase), EC 3.1.4.12 (Sphingomyelin phosphodiesterase), EC 4.2.99.18 (DNA-(apurinic or apyrimidinic site) lyase), and EC 5.3.99.3 (Prostaglandin-E synthase).

Different Biological Roles or Functional Redundancy?

In this work, we found substantial evidence of nonhomologous isofunctional enzymes coexisting in 15 enzymatic activities (comprising 70 enzymatic sequences) of human metabolism. Notably, despite the use of very restrictive criteria (excluding multimeric enzymes, enzymatic activities with incomplete EC classification, as well as clusters composed exclusively of a single human sequence) and our focus on human enzymatic activities in which the participation of unrelated enzymes are recognized, we discovered intragenomic analogous enzymes in three enzymatic activities (20% of our bona fide data set) with no evidence of analogy reported so far: lactose synthase (EC 2.4.1.22), NAD+—protein-arginine ADP-ribosyltransferase (EC 2.4.2.31), and 1-phosphatidylinositol 4-kinase (EC 2.7.1.67). These enzymatic activities participate in nine distinct biochemical pathways or biological processes, some of which playing essential roles in cancer, galactose metabolism, glycosaminoglycan biosynthesis, glycosphingolipid biosynthesis, inositol phosphate metabolism, mannose type O-glycan biosynthesis, n-glycan biosynthesis, other types of o-glycan biosynthesis, and phosphatidylinositol signaling system.

We hypothesize that the coexistence of multiple nonhomologous isofunctional enzymes in the human metabolism might not be interpreted as functional redundancy since these intragenomic analogous enzymes might be implicated in distinct biological roles. To test this hypothesis, we will be comparing the transcription profile of the genes encoding the repertoire of intragenomic analogous enzymes cataloged in human metabolism, using RNA-Seq data obtained from 8,555 samples of 53 distinct healthy human tissues publicly available at the GTEx portal (The GTEx Consortium 2013) (https://www.gtexportal.org/home/). The identification of alternative enzymatic forms differentially expressed or coexpressed could provide evidence regarding possible distinct biological roles played by human intragenomic analogous enzymes.

Materials and Methods

Computational Prediction of Analogy

Protein sequences from 2,494 completely sequenced genomes comprising organisms of the three domains of life were obtained from the KEGG database release 73.1 (Kanehisa and Goto 2000) (http://www.genome.jp/kegg/) and clustered by enzymatic activity, based on the degree of similarity between their amino acid sequences, applying the methodology described in Otto et al. (2008), implemented in AnEnPi pipeline; sequences sharing the same

enzymatic activity but assigned to two or more distinct clusters are considered putative functional analogous, indicating one or more possible events of independent evolutionary origin.

Briefly, we compared 1,159,633 enzymatic sequences, separately by enzymatic activity (EC), all against all, using BLAST+ version 2.2.30 (Altschul 1997) and default parameters. Next, we transformed the sequence alignment result in a graph in which each enzymatic sequence represents a node. For each enzymatic activity, any sequence (node) pair that achieved an alignment score ≥ 120 were connected by an edge; linked sequences are presumably homologous, therefore were grouped in the same AnEnPi cluster; on the other hand, enzymatic sequences grouped in distinct AnEnPi clusters of the same enzymatic activity are presumably analogous. The number of subgraphs obtained represents the number of putative events of independent origin in each enzymatic activity or, in other words, the number of times a particular enzymatic activity has arisen during evolution. The similarity threshold used in the clustering phase (BLAST score ≥ 120) is based on a significant experimental observation: enzymes that proved to share the same enzymatic activity and present significantly different 3D structures (based on structural alignments), scored below 120 when their amino acid sequences were compared with BLAST (Galperin et al. 1998). Although the absence of detectable sequence similarity might often be attributed to the divergence between homologous sequences during evolution, it was observed that many alternative forms of enzymes catalyzing the same biochemical reaction had significantly distinct 3D structures, and therefore have (presumably) evolved independently (Galperin et al. 1998; Omelchenko et al. 2010).

Subsequently, the AnEnPi output was processed as follows: 1) incomplete ECs were removed. Enzymes whose chemical transformations are defined only up to the third digit of the EC classification may have different reaction specificities (different substrates/products or cofactors) and, in this case, the predicted analogues would correspond to a mechanistic analogy. However, this type of analogy is not part of the scope of this work, which is devoted exclusively to the study of functional analogy; 2) enzymatic activities in which enzymes were annotated as "subunit" and "chain" were manually inspected and excluded, because the presence of heteromultimeric enzymes in the data set can inflate the number of analogy events detected. This problem arises during the process of annotation of enzymatic sequences, in which different subunits (or chains) of a multimeric enzyme often inherit the annotated activity for the enzyme as a whole disregarding its evolutionary origin and its participation in the related activity; 3) enzymatic activities containing clusters composed exclusively of a single human sequence were removed. If a single sequence distinguishes from tens or hundreds of other enzymatic sequences (from humans and/or other species) we consider it suspicious, as this might represent functional misannotation; 4) enzymatic activities in which the occurrence of alternative enzymatic forms was not detected (composed of a single cluster) were removed.

Validation of Predicted Intragenomic Analogy in Human

We used protein domains, superfamily/folding, and 3D structure annotations retrieved for proteins within and between clusters of each enzymatic activity to confirm putative cases of analogy detected inside the human genome (intragenomic analogy). These data were obtained from Pfam 27.0 (Finn et al. 2014) (http://pfam.xfam.org/) and SUPERFAMILY 1.75 (Wilson et al. 2009) (http://supfam.org/SUPERFAMILY/).

Experimentally resolved 3D structures for proteins were retrieved from PDB database (Berman et al. 2000) (http://www.rcsb.org/) and previous information of convergence in enzymatic activities were selected from the scientific literature (Omelchenko et al. 2010). Based on superfamily classification, we split our data set into two data sets: (−) and (+) data set. Enzymatic activities composed of putative alternative enzymatic forms belonging to at least two distinct superfamilies were assigned to the (+) data set, otherwise were assigned to the (−) data set.

For the bona fide (+) data set, we performed an all-against-all pairwise sequence comparison among all sequences inside each enzymatic activity using the optimal global sequence alignment implemented in the software Needle (Rice et al. 2000).

We generated structural models for sequences without 3D information employing the comparative modeling software Modeller (Webb and Sali 2014). For this, we used templates from PDB database retrieved by BLAST similarity searches (coverage in query $> 70\%$; coverage in subject $> 90\%$; identity $> 30\%$; *e*-value $< 10\text{-}3$). Modeller generated 50 structural models and the best model for each protein was selected based on the lowest DOPE (Discrete Optimized Protein Energy) score value. Subsequently, the quality of these selected models were evaluated with SAVES (http://services.mbi.ucla.edu/SAVES/) and MolProbity (Chen et al. 2010). The side chains were fitted with KiNG (Chen et al. 2009), and the energy minimization was performed with ModRefiner (Xu and Zhang 2011).

The 3D structures were generated with PyMOL (The PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC) and the protein structural alignments were performed with TM-align (Zhang and Skolnick 2005). RMSD values and TM-scores (Zhang and Skolnick 2004) were calculated with the TM-align package. TM-score distances were normalized by the average size of the chains of each compared structure.

Genomic and Metabolic Mapping of Analogous Enzymes

Genomic coordinates of genes encoding the alternative forms in the bona fide (+) data set were retrieved from Ensembl

(genome version: GRCh38) (Cunningham et al. 2015). The ideogram representing the chromosomal localization of these genes was created with the software PhenoGram (http://ritchielab.psu.edu/). Additionally, a circular diagram displaying the genomic distances between genes encoding alternative forms (distinct AnEnPi cluster of the same EC) as well as genes encoding homologous enzymatic forms (belonging to the same AnEnPi cluster and EC) was created with Circos (Krzywinski et al. 2009).

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We wish to thank Fernando Alvarez-Valín, Marcelo Alves Ferreira, and Leandro Mattos Pereira for fruitful discussions; André Luiz Quintanilha Torres and Luiz Phillippe Ribeiro Baptista for their help in the analysis. We also thank to Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and the Plataforma de Bioinformática da Fiocruz RPT04A/RJ for their support.

Literature Cited

Almonacid DE, Yera ER, Mitchell JBO, Babbitt PC. 2010. Quantitative comparison of catalytic mechanisms and overall reactions in convergently evolved enzymes: implications for classification of enzyme function. PLoS Comput Biol. 6:e1000700.

Altschul S. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Arimitsu E, et al. 1999. Cloning and sequencing of the cDNA species for mammalian dimeric dihydrodiol dehydrogenases. Biochem J. 342:721.

Barylko B, et al. 2001. A novel family of phosphatidylinositol 4-kinases conserved from yeast to humans. J Biol Chem. 276:7705–7708.

Berman HM, et al. 2000. The Protein Data Bank and the challenge of structural genomics. Nat Struct Biol. 7:957–959.

Boura E, Nencka R. 2015. Phosphatidylinositol 4-kinases: function, structure, and inhibition. Exp Cell Res. 337:136–145.

Bridges PJ, et al. 2012. Hematopoetic prostaglandin D synthase: an ESR1-dependent oviductal epithelial cell synthase. Endocrinology 153:1925–1935.

Capriles PVSZ, et al. 2010. Structural modelling and comparative analysis of homologous, analogous and specific proteins from *Trypanosoma cruzi* versus *Homo sapiens*: putative drug targets for chagas' disease treatment. BMC Genomics 11:610.

Carbone V, Hara A, El-Kabbani O. 2008. Structural and functional features of dimeric dihydrodiol dehydrogenase. Cell Mol Life Sci. 65:1464–1474.

Chen VB, et al. 2010. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr. 66:12–21.

Chen VB, Davis IW, Richardson DC. 2009. KING (Kinemage, Next Generation): a versatile interactive molecular and scientific visualization program. Protein Sci. 18:2403–2409.

Cordwell SJ. 1999. Microbial genomes and "missing" enzymes: redefining biochemical pathways. Arch Microbiol. 172:269–279.

Crisp A, Boschetti C, Perry M, Tunnacliffe A, Micklem G. 2015. Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. Genome Biol. 16:50.

Cunningham F, et al. 2015. Ensembl 2015. Nucleic Acids Res. 43:D662–D669.

Doolittle RF. 1994. Convergent evolution: the need to be explicit. Trends Biochem Sci. 19:15–18.

Finn RD, et al. 2014. Pfam: the protein families database. Nucleic Acids Res. 42:D222–D230.

Galperin MY, Koonin EV. 1999. Functional genomics and enzyme evolution. Homologous and analogous enzymes encoded in microbial genomes. Genetica 106:159–170.

Galperin MY, Koonin EV. 2012. Divergence and convergence in enzyme evolution. J Biol Chem. 287:21–28.

Galperin MY, Walker DR, Koonin EV. 1998. Analogous enzymes: independent inventions in enzyme evolution. Genome Res. 8:779–790.

Garcia YM, et al. 2017. A superoxide dismutase capable of functioning with iron or manganese promotes the resistance of *Staphylococcus aureus* to calprotectin and nutritional immunity. PLOS Pathog. 13:e1006125.

George RA, Spriggs RV, Thornton JM, Al-Lazikani B, Swindells MB. 2004. SCOPEC: a database of protein catalytic domains. Bioinformatics 20:i130–i136.

Gherardini PF, Wass MN, Helmer-Citterich M, Sternberg MJE. 2007. Convergent evolution of enzyme active sites is not a rare phenomenon. J Mol Biol. 372:817–845.

Hanson AD, Pribat A, Waller JC, Crécy-Lagard V. d. 2010. "Unknown" proteins and "orphan" enzymes: the missing half of the engineering parts list: and how to find it. Biochem J. 425:1–11.

Hegyi H, Gerstein M. 1999. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. J Mol Biol. 288:147–164.

Heilmeyer LMG, Vereb G, Vereb G, Kakuk A, Szivák I. 2003. Mammalian phosphatidylinositol 4-kinases. IUBMB Life 55:59–65.

Huynen M. a, Dandekar T, Bork P. 1999. Variation and evolution of the citric-acid cycle: a genomic perspective. Trends Microbiol. 7:281–291.

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28:27–30.

Krzywinski M, et al. 2009. Circos: an information aesthetic for comparative genomics. Genome Res. 19:1639–1645.

Landis GN, Tower J. 2005. Superoxide dismutase evolution and life span regulation. Mech Ageing Dev. 126:365–379.

Lim SM, et al. 2013. Structural and dynamic insights into substrate binding and catalysis of human lipocalin prostaglandin D synthase. J Lipid Res. 54:1630–1643.

Marín-Méndez JJ, et al. 2012. Differential expression of prostaglandin D2 synthase (PTGDS) in patients with attention deficit-hyperactivity disorder and bipolar disorder. J Affect Disord. 138:479–484.

Martin W, Schnarrenberger C. 1997. The evolution of the Calvin cycle from prokaryotic to eukaryotic chromosomes: a case study of functional redundancy in ancient pathways through endosymbiosis. Curr Genet. 32:1–18.

McDonald AG, Tipton KF. 2014. Fifty-five years of enzyme classification: advances and difficulties. FEBS J. 281:583–592.

Morett E, et al. 2003. Systematic discovery of analogous enzymes in thiamin biosynthesis. Nat Biotechnol. 21:790–795.

Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol. 247:536–540.

Omelchenko MV, Galperin MY, Wolf YI, Koonin EV. 2010. Non-homologous isofunctional enzymes: a systematic analysis of alternative solutions in enzyme evolution. Biol Direct. 5:31.

Otto TD, Guimarães ACR, Degrave WM, de Miranda AB. 2008. AnEnPi: identification and annotation of analogous enzymes. BMC Bioinformatics 9:544.
Penning TM, Chen M, Jin Y. 2015. Promiscuity and diversity in 3-ketosteroid reductases. J Steroid Biochem Mol Biol. 151:93–101.
Peregrin-Alvarez JM, Tsoka S, Ouzounis CA. 2003. The phylogenetic extent of metabolic enzymes and pathways. Genome Res. 13:422–427.
Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European molecular biology open software suite. Trends Genet. 16:276–277.
Sillitoe I, et al. 2015. CATH: comprehensive structural and functional annotations for genome sequences. Nucleic Acids Res. 43:D376–D381.
Tanaka K, et al. 2000. Cutting edge: differential production of prostaglandin D2 by human helper T cell subsets. J Immunol. 164:2277–2280.
The GTEx Consortium. 2013. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 45:580–585.
Trimarco A, et al. 2014. Prostaglandin D2 synthase/GPR44: a signaling axis in PNS myelination. Nat Neurosci. 17:1682–1692.
Urade Y, Eguchi N. 2002. Lipocalin-type and hematopoietic prostaglandin D synthases as a novel example of functional convergence. Prostaglandins Other Lipid Mediat. 68–69:375–382.
Webb B, Sali A. 2014. Comparative protein structure modeling using MODELLER. Curr Protoc Bioinform. 47:5.6.1-32.
Wilson D, et al. 2009. SUPERFAMILY: sophisticated comparative genomics, data mining, visualization and phylogeny. Nucleic Acids Res. 37:D380–D386.
Xu D, Zhang Y. 2011. Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. Biophys J. 101:2525–2534.
Zhang Y, Skolnick J. 2004. Scoring function for automated assessment of protein structure template quality. Proteins 57:702–710.
Zhang Y, Skolnick J. 2005. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 33:2302–2309.
Zimmermann H. 1992. 5′-Nucleotidase: molecular structure and functional aspects. Biochem J. 285:345–365.
Zukowska P, Kutryb-Zajac B, Toczek M, Smolenski RT, Slominska EM. 2015. The role of ecto-5′-nucleotidase in endothelial dysfunction and vascular pathologies. Pharmacol Rep. 67:675–681.

Associate editor: Bill F. Martin