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.

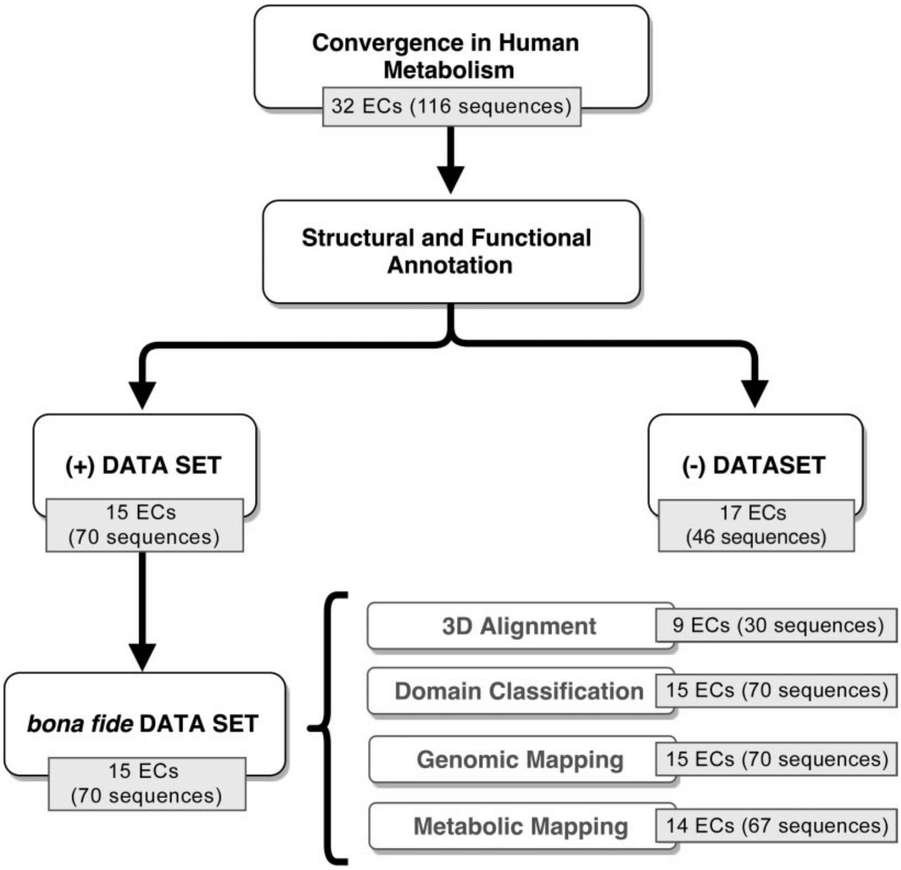

FIG. 1.—Outline of the procedure used for the identification of intragenomic analogy in human metabolism (see text for details).

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

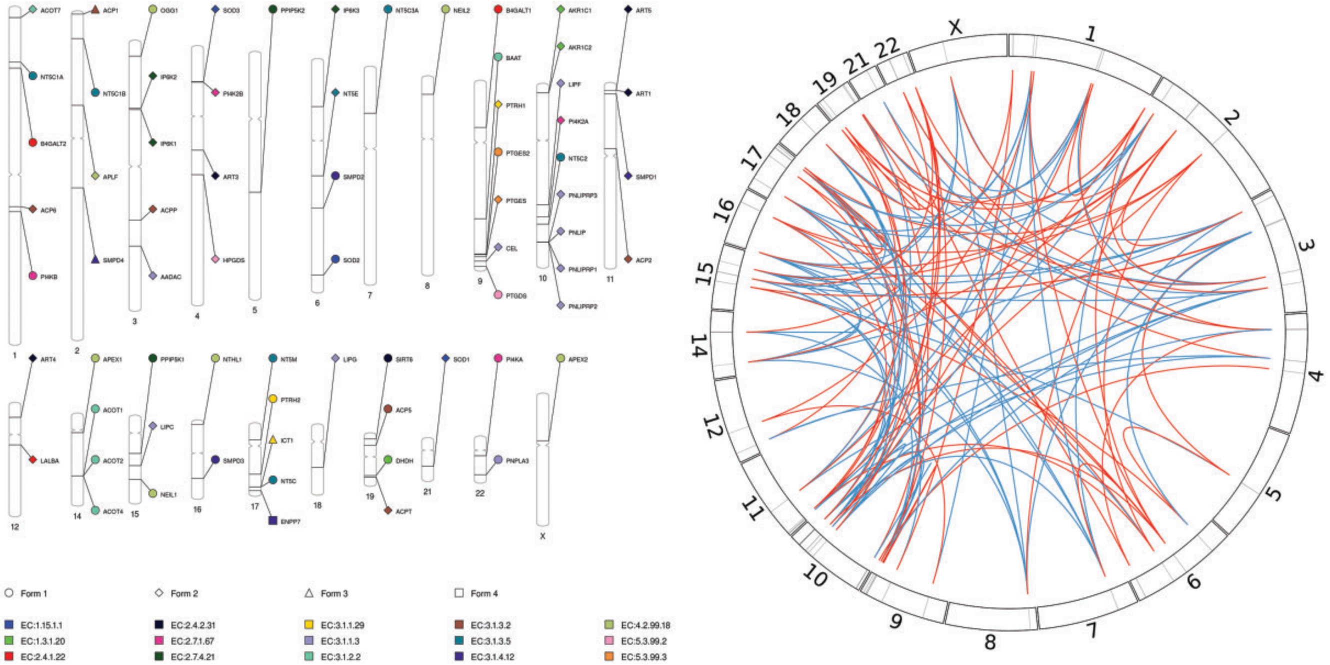

Fig. 2.—(Left) Diagram depicting the localization of genes encoding intragenomic analogous enzymes across the human chromosome. Enzymatic activities in which evidence of intragenomic analogy was found are represented by distinct colors. Genes encoding distinct enzymatic forms are represented by different 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 < 10-3). 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