Bioinformation Volume 5 Issue 4

www.bioinformation.net

open access

Hypothesis

Classifying glycerol dehydratase by its functional residues and purifying selection in its evolution

Andres Julian Gutierrez Escobar[1]*, Dolly Montoya Castaño[1]

[1]Universidad Nacional de Colombia, Bioprocesses and Bioprospecting Research Group, Instituto de Biotecnología Bogotá, Colombia; Andres Julian Gutierrez Escobar- Email: andresgutierrez@colombia.com; *Corresponding author.

Received July 25, 2010; Accepted August 25, 2010; Published October 06, 2010

Abstract:
Glycerol dehydratase (GD) catalyses glycerol reductive conversion to 3-hydroxypropanaldehyde (3-HPA), this being the first step required for the microbial conversion of glycerol to 1, 3 -propanodiol. GD has been functionally characterised to date and two main groups have been determined, one of them being vitamin B_{12}-dependent and the other B_{12}-independent. GD evolutionary history has been described and an exhaustive analysis made for detecting the functional residues responsible for type I divergence. GD phylogenetic tree topology was seen to be statistically robust and the data indicated strong purifying selection operating on the GD proteins within it. Two clades were indentified, one for vitamin B_{12}-dependent and the other for B_{12}-independent classes. The ancient hot-pot residues responsible for protein divergency for each clade were also identified. The basic evolutionary biology for GD proteins has been described, thereby opening the way forward for developing rational mutagenesis studies.

Key words: glycerol dehydratase, molecular evolution, type I functional divergence, hot pots.

Background:

Interest in glycerol dehydratase GD (EC 4.2.1.30) has increased beyond academic circles in the past few years because of its role in the fermentation pathway for producing industrial 1,3- propanediol (1,3-PD). Two kinds of GD have been characterised to date. The first one catalyses glycerol conversion to 3-hydroxypropanaldehyde via a radical mechanism depending on the extensively studied 5'-deoxyadenosylcobalamin (vitamin coB_{12}) **[1]**; the other performs the same function but is B_{12}-independent. Both enzymes belong to the new radical SAM superfamily of proteins which has been identified in all kingdoms of life and has been shown to catalyse a diverse array of chemical reactions having significant medical and biotechnological importance. The GDs specifically belongs to the lyase family which cleaves carbon-oxygen bonds **[2]**.

The cofactors required for such common activation mechanism are a $[4Fe\text{-}4S]^+$ cluster (three Fe2+ ions and one Fe3+ ion) and S-adenosylmethionine (SAM). Glycerol dehydratase is a key enzyme for the dihydroxyacetone (DHA) pathway **[3]**. The C. butyricum enzyme presents the highest identity (47%) with E. coli PFL (piruvate formate lyase) according to Raynaud et al., specifically the C-terminal domain (the radical loop). Its overall structure is an β/α barrel containing its catalytic properties. The B_{12}-independent enzyme forms a monomer forming a functional dimer **[4]**; however, the B_{12}-dependent one exists as an αβγ heterotrimer dimer. The α monomer corresponds to the β/α barrel **[5]**.

Neither the basic evolutionary biology for this class of protein nor the type of residues considered to be evolutionary hot spots has been deduced at the present. This study has examined GD molecular evolutionary history to determine whether the evolutionary process has been responsible for the high degree of sequence conservation. Different methodological approaches were used for analysing synonymous (pS) and non-synonymous (pN) changes in 31 GD sequences. PRATT software was used for predicting the GD motif signature and the Evolutionary Trace server was used for determining evolutionary traces for the GD protein. Specific amino acids responsible for selective restriction were then identified, phylogenetic divergence being produced for this protein. DIVERGE 1.0 software was used in our approach for evaluating all protein sequences.

Methodology:

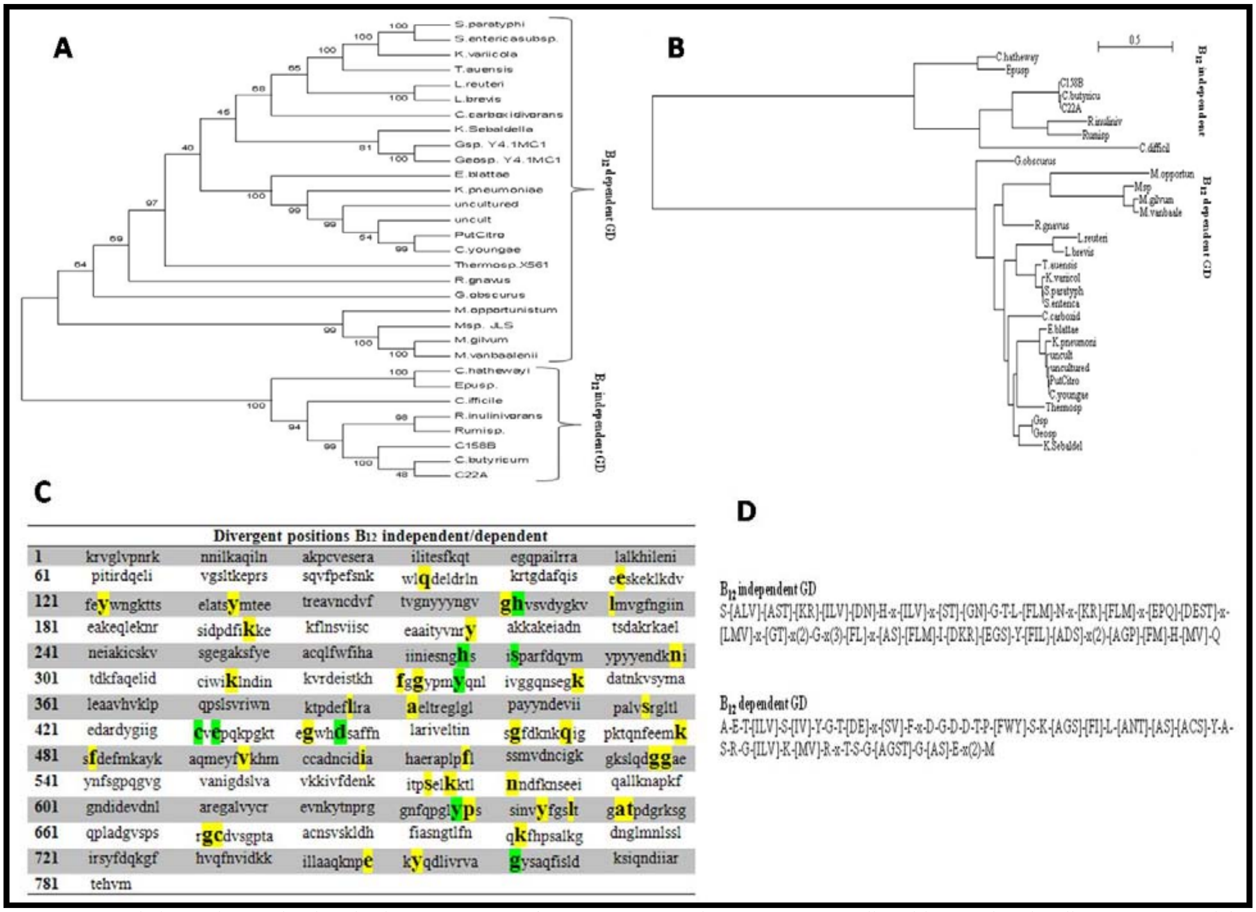
Sequences:

An exhaustive search was made in GenBank, EMBL and Swiss-prot databases for GD nucleotide and protein sequences. This search was optimised by using BLAST, PSI-BLAST and WU-BLAST software **(6)** using the Clostridium butyricum protein sequence as search entry (access number ABX56860.2). 103 hits were obtained and then filtered by removing partial and redundant sequences from the population. Complete protein representations were included by strain; our final working population consisted of 31 complete protein sequences. SMART software was used for scrutinising all sequences in the search for typical GD protein domains **[7]**. GD crystal structures were downloaded from the PDB database; the 1r9d structure **[4]** was used as template for divergent functional residue analysis.

Alignment and phylogenetic reconstruction:

Muscle software **[8]** was used for gene and protein alignment of the 31 previously collected sequences, using default parameters. dS and dN percentage changes were computed using a modified version of the Nei-Gojobori test; the Tajima test was calculated using MEGA 4.0 software and the SNAP server **[9]**. A combined strategy was used for phylogenetic analysis; the NJ method was used first for phylogenetic reconstruction and p-distance as a model for distance analysis **[10]**. Statistical robustness was calculated by using 5,000 Bootstrap repeats. MEGA 4.0 software was used throughout **[11]**.

Secondly, the alignment was then analysed using ProtTest **[12]** to determine the protein evolution model having the best fit for GD sequence alignment. Phylogenetic analysis then used Phyml 3.0.1 **[13]**, using 1,000 Bootstrap repeats. The phylogenetic tree was then visualised using NJplot software **[14]**. The best tree topology was shown.



Divergent positions B12 independent/dependent

1	krvglvpnrk	nnilkaqiln	akpcvesera	ilitesfkqt	egqpailrra	lalkhileni
61	pitirdqeli	vgsltkeprs	sqvfpefsnk	wlqdeldrln	krtgdafqis	eeskeklkdv
121	feywngktts	elatsymtee	treavncdvf	tvgnyyyngv	ghvsvdygkv	lmvgfngiin
181	eakeqleknr	sidpdfikke	kflnsviisc	eaaityvnry	akkakeiadn	tsdakrkael
241	neiakicskv	sgegaksfye	acqlfwfiha	iiniesnghs	isparfdqym	ypyyendkni
301	tdkfaqelid	ciwiklndin	kvrdeistkh	fggypmyqnl	ivggqnsegk	datnkvsyma
361	leaavhvklp	qpslsvriwn	ktpdefllra	aeltreglgl	payyndevii	palvsrgltl
421	edardygiig	cvepqkpgkt	egwhdsaffn	lariveltin	sgfdknkqig	pktqnfeemk
481	sfdefmkayk	aqmeyfvkhm	ccadncidia	haeraplpfl	ssmvdncigk	gkslqdggae
541	ynfsgpqgvg	vanigdslva	vkkivfdenk	itpselkktl	nndfknseei	qallknapkf
601	gndidevdnl	aregalvycr	evnkytnprg	gnfqpglyps	sinvyfgslt	gatpdgrksg
661	qpladgvsps	rgcdvsgpta	acnsvskldh	fiasngtlfn	qkfhpsalkg	dnglmnlssl
721	irsyfdqkgf	hvqfnvidkk	illaaqknpe	kyqdlivrva	gysaqfisld	ksiqndiiar
781	tehvm					

B_{12} independent GD
S-[ALV]-[AST]-[KR]-[ILV]-[DN]-H-x-[ILV]-x-[ST]-[GN]-G-T-L-[FLM]-N-x-[KR]-[FLM]-x-[EPQ]-[DEST]-x-[LMV]-x-[GT]-x(2)-G-x(3)-[FL]-x-[AS]-[FLM]-I-[DKR]-[EGS]-Y-[FIL]-[ADS]-x(2)-[AGP]-[FM]-H-[MV]-Q

B_{12} dependent GD
A-E-T-[ILV]-S-[IV]-Y-G-T-[DE]-x-[SV]-F-x-D-G-D-D-T-P-[FWY]-S-K-[AGS]-[FI]-L-[ANT]-[AS]-[ACS]-Y-A-S-R-G-[ILV]-K-[MV]-R-x-T-S-G-[AGST]-G-[AS]-E-x(2)-M

Figure 1: (A) Phylogenetic tree developed using MEGA 4.0 software, 31 GD protein sequences were aligned in Muscle. The alignment was used for constructing a tree using the NJ, p-distance and 5,000 Bootstrap repetitions for statistical robustness **(B)** Phylogenetic tree developed using Phyml software, 31 GD protein sequences were aligned in Muscle. The alignment was used for constructing a tree using the JTT + γ evolutionary model according results from Prottest, 5,000 Bootstrap repetitions was used for statistical robustness. Only nodes having values higher than 50% statistical significance have been shown (refer to supporting material for phylip sequence format used) C) Determining functionally important phylogenetically divergent sites for the GD protein. The sites have a θ-value of 0.7 for all 39 sites (yellow), these being statistically significant values (falling within the 5% region, having a P-value <0.05). Green residues considered the enzymatic active site according literature. D) PRATT traces residues for B_{12} Independent and dependent types.

Analysing type I functional residues:

A conceptual statistical framework for modelling functional divergence was used for estimating the coefficient of functional divergence **(θ)** as type I functional divergence level indicator. GD protein alignments were used for determining divergence points (DIVERGE software 1.0) **[16]**.

Discussion:

Glycerol conversion to 1, 3-PD involves a B_{12}-dependent glycerol dehydratase coenzyme **[5]**. However, one report has described that Clostridium butyricum VPI1718 glycerol dehydratase (extracted from 1,3-PD-producing cells) was not stimulated by coenzyme B_{12} and was extremely oxygen sensitive, thereby suggesting that it might be a B_{12}-independent coenzyme **[4]**. It seems that B_{12}-dependent and B_{12}-independent enzymes are orthologous genes which have evolved in separate lines; however, β/α barrel homology indicate an ancestral relationship.

GD evolution is characterised by ancient gene duplications (supported by high basal bootstrap values) followed by bifurcation having long branches, indicating independent evolution for each clade. Despite similar tree branching being observed when using both strategies (see methodology), the second one seemed to be the most parsimonious because it required less steps to reproduce the topology with a good bootstrap value **(Figure 1A, B)**. Interestingly, the longer basal branches of the tree (1,756 for B_{12}-independent and 2,175 for B_{12}-dependent nodes) indicated a deep common ancestor even though each current GD clade has its own evolutionary mode. This hypothesis has been demonstrated by structural analysis for both enzymatic types in which the B_{12}-dependent type has additional chains (contrary to the B_{12}-independent types). JTT+γ was the evolutionary model which best fit our protein sequences **[17]**; this was not calculated by MEGA 4.0 but is default in Phyml 1.0 software. This strategy has been seen to be effective in predicting the best model for GD evolution.

Several approaches were applied for testing natural selection. The results suggested that dS level was higher than dN **(Table 1 see supplementary material)**. A 0.000 probability was obtained in the Z-test (dS–dN=3.538). Tajima D value was 4.857740 and dS/dN was 1.3432 in the SNAP server. This suggested that birth and death subjected to strong purifying selection was the model best fitting GD protein evolution.

Such combination has thus sought the best polymorphism by niche, explored according to species. This indicated that GD genes have been in the bacterial genome for a long time. It also suggested that GD was a determinant point of natural selection and thereby cooperated by inducing the divergence of these kinds of bacterial species. It is possible that the GD protein belongs to the radical SAM superfamily but the blast result suggested that it fit better with the RNR-PFL superfamily (data not shown). Such enzymes are strictly anaerobic (like GD) and it has been further suggested that the diversity of chemical reactions catalysed by this class of protein exceeds those catalysed by B_{12} **[18]**. Glycerol is the primary metabolite of GD but has a wide variety of catalysed substrates according to its evolutionary mode. GD displayed broad spectrum substrates in this work. GD can catalyse 1, 2-ethanediol → acetaldehyde + H_2O, 1,2-propanediol → propionaldehyde + H_2O **[19]** and ethylene glycol → acetaldehyde + H_2O (20) and GD may have a plethora of substrates which have not yet been discovered.

Several residues have been determined for GD function. GD has been found at Gly763 within the Clostridia Gly-radical domain (which has been identified as being the site for free radical formation) and Cys433 located around it. The active site binding glycerol and 1, 2-propanediol are mediated by H281, H164, S282, D447, E435, Y640, C433 and Y339 residues (1–4). R782 may be important for functional contact between GD and its reactivase protein **[4]**.

The cut-off value for detecting type I divergent residues was 0.7. Thirty-nine residues were detected here (Q94, E203, Y124, Y137, G162, L172, k199, Y212, N100, K316, F332, G33, K350, L388, A392, S416, G443, G463, Q469, K481, F483,Y498, I510, F520, G539, G538, S575, K578,N582, P640, Y646, L650, A653, T654, G672, C673, K703, E751 and Y753) which can be considered to be hot-spots for GD evolution. These sites may be the mutational points defining B_{12}-dependent and B_{12}-independent GD lines (for more details please refer to supporting material). PRATT was used for obtaining the trace maps for each GD protein **(Figure 1D)**. Some of these sites fit just at the side of functionally proven residues from the GD active site (i.e. G162/H163 and Y639/P640). Such changes protect protein function but generate protein distortions opening up the sequence space for exploring new niches.

Conclusion:

For one hand, GD protein evolution can be clearly explained by birth and death evolution in purifying selection mode and opens the way forward for future mutagenesis studies pursuing enzymatic activity improvement based on the traces identified here. For the other hand, it is important to develop non conventional data mining strategies looking for the optimal identification of RNF-PFL proteins family members in the databases.

Acknowledgement:

We would like to thank the Bioprocesses and Bioprospecting research group from the Biotechnology Institute at the National University of Colombia and also the ACAC (Asociación Colombiana para el avance de la ciencia) for their support in this paper. Finally, we give thanks to Alfonso Benítez Páez for his suggestions about evolutionary topics and Jason Garry for his contribution towards reviewing the text.

References:

[1] C Raynaud *et al. PNAS* **9**: 5010 (2003) [PMID: 12704244]
[2] H Sofia *et al. Nucleic Acids Res* **29**: 1097 (2001) [PMID: 11222759]
[3] R Forage & E Lin. *J. Bacteriol* **5**: 91 (1982) [PMID: 7035429]
[4] J O'Brien *et al. Biochemistry* **16**: 4635 (2004) [PMID: 15096031]
[5] M Yamanishi *et al. Eur. J. Biochem*. **18**: 4484 (2002) [PMID: 12230560]
[6] Altschul *et al. Nucleic Acids Res* **17**:3389 (1997) [PMID: 9254694]
[7] I Letunic *et al. Nucleic Acids Res* **37**: 229 (2009) [PMID: 18978020]
[8] R Edgar. *Nucleic Acids Res* **32**: 1792 (2004) [PMID: 15318951]
[9] www.hiv.lanl.gov/content/sequence/SNAP/SNAP.html
[10] M Kimura. *J Mol Evol* **16**: 111 (1980) PMID: [7463489]
[11] S Kumar *et al. Brief Bioinform* **4**: 299 (2008) [PMID: 18417537]
[12] F Abascal *et al. Bioinformatics* **9**:2104 (2005) [PMID: 15647292]
[13] S Guindon & O Gascuel. *Syst Biol* **5**:696 (2003) [PMID: 14530136]
[14] G Perrière & M Gouy. *Biochemis* **78**: 364 (1996) [PMID: 8905155]
[15] I Jonassen *et al. Prot Scien* **8**:1587 (1995) [PMID: 8520485]
[16] X Gu. *Mol Biol Evol* **12**:1664 (1999) PMID: [10605109]
[17] DT Jones *et al., Comput Appl Bioscn* **8**: 275 (1992) [PMID: 1633570]
[18] PA Frey. *Chem. Rev* **103**: 2129 (2003) [PMID: 12797826]
[19] T Toraya & S Fukui. *Eur J Biochem* **1**:285 (1977) [PMID: 407082]
[20] A Stroinski *et al. Arch Biochem Biophys* **2**:321 (1974) [PMID: 4407304]

Edited by P. Kangueane
Citation: Escobar *et al.* Bioinformation 5(4): 173-176 (2010)

Supplementary material:

Table 1: Observed evolutionary properties of GD proteins.

Test	Parameters					
Tajima	M	S	Ps	θ	Π	D
	31	517	0.982890	0.246031	0.553898	4.857740
Z-test	Neutral/stat	Positive/Stat	Purifying/stat			
	0.001/-3.558	1.000/-3.558	0.000/3.558			
SNAP	dS	dN	dS/dN	ps/pn		
	1.3451	0.9895	1.3432	1.1186		

The Tajima test statistic was estimated using MEGA4. All positions containing gaps and missing data were eliminated from the dataset (Complete deletion option). The abbreviations used are as follows: m = number of sites, S = Number of segregating sites, p_s = S/m, $\Theta = p_s/a_1$ and π = nucleotide diversity. D is the Tajima test statistic (see chapter 12 in ref. 3 for equations). The Z-test was estimated in overall population using MEGA4. All positions containing gaps and missing data were eliminated from the dataset. The SNAP server was using with standard bacterial genetic code using nucleotide sequences two sequences were eliminates for the populations.