Journal of Theoretical Biology 357 (2014) 160–168

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Prediction of the determinants of thermal stability by linear discriminant analysis: The case of the glutamate dehydrogenase protein family

Angelo Pavesi *

Department of Life Sciences, University of Parma, Parco Area delle Scienze 11/A, I-43124 Parma, Italy

HIGHLIGHTS

- A novel method for predicting the determinants of thermostability in a protein family.
- It is a multivariate statistical method based on the linear discriminant analysis.
- It was applied to a set of 69 glutamate dehydrogenases from Archaea and Bacteria.
- Three amino acid clusters were predicted to be the determinants of thermostability.
- Analysis within the clusters led to identification of 8 critical amino acid sites.

ARTICLE INFO

Article history:
Received 3 December 2013
Received in revised form
7 May 2014
Accepted 8 May 2014
Available online 20 May 2014

Keywords:
Archaea
Bacteria
Mesophile
Multivariate statistics
Thermophile

ABSTRACT

Little is known about the determinants of thermal stability in individual protein families. Most of the knowledge on thermostability comes, in fact, from comparative analyses between large, and heterogeneous, sets of thermo- and mesophilic proteins. Here, we present a multivariate statistical approach aimed to detect signature sequences for thermostability in a single protein family. It was applied to the glutamate dehydrogenase (GDH) family, which is a good model for investigating this peculiar process. The structure of GDH consists of six subunits, each of them organized into two domains. Formation of ion-pair networks on the surface of the protein subunits, or increase in the inter-subunit hydrophobic interactions, have been suggested as important factors for explaining stability at high temperatures. However, identification of the amino acid changes that are involved in this process still remains elusive. Our approach consisted of a linear discriminant analysis on a set of GDH sequences from Archaea and Bacteria (33 thermo- and 36 mesophilic GDHs). It led to detection of 3 amino acid clusters as the putative determinants of thermal stability. They were localized at the subunit interface or in close proximity to the binding site of the $NAD(P)^+$ coenzyme. Analysis within the clusters led to prediction of 8 critical amino acid sites. This approach could have a wide utility, in the ligth of the notion that each protein family seems to adopt its own strategy for achieving thermostability.

1. Introduction

Glutamate dehydrogenase (GDH) catalyzes the reversible oxidative deamination of L-glutamate to 2-oxoglutarate and ammonia using $NAD(P)^+$ as coenzyme. The structure of GDH consists of six identical subunits of 48 kDa, each of them organized into two domains separated by a deep cleft. Domain I is responsible for directing the self-assembly of the subunits into hexamer, while domain II forms the dinucleotide-binding domain (Baker et al., 1992).

The broad characterization of GDH in prokaryotes has led to its adoption as a paradigm for investigating adaptation to temperatures around 100 °C. Comparison between the 3D structure of the thermophilic GDH from *Pyrococcus furiosus* and that of the mesophilic homolog from *Clostridium symbiosum* has revealed one major difference, consisting of a series of ion-pair networks on the surface of the protein subunits and buried at inter-domain and inter-subunit interfaces (Yip et al., 1995). This finding has been supported by mutational studies on the GDHs of *Thermococcus litoralis* (Vetriani et al., 1998) and *Thermotoga maritima* (Lebbink et al., 1999). Determination of the 3D structure of the hyperthermostable GDH from *Pyrobaculum islandicum* has revealed that oligomerization is also affected by an increase in the inter-subunit hydrophobic interactions (Bhuiya et al., 2005).

* Tel.: +390521905647; fax: +390521905604.
E-mail address: angelo.pavesi@unipr.it

http://dx.doi.org/10.1016/j.jtbi.2014.05.013
0022-5193/

Analysis of the amino acid composition has shown that thermophilic GDHs have a lower content of glycine, methionine, and cysteine with respect to mesophilic GDHs (Knapp et al., 1997). Bhuiya et al. (2005) have found that the hyperthermostable GDHs of *P. islandicum* and *P. furiosus* show a similar content of alanine and β−branched residues (Val, Thr, and Ile). Comparison of the mesophilic GDH from *C. symbiosum* with the thermophilic homologs from *P. furiosus* and *T. litoralis* has pointed out that the most frequent substitutions are Val/Ile, Val/Lys, Leu/Ile, and Gly/Ala, though no solid conclusions about their functional role have been reached (Britton et al., 1995).

A criticism to this approach is that discrimination between the amino acid changes responsible for thermal adaptation and those affected by other evolutionary pressures is very difficult to achieve. For example, the GDH of *P. furiosus* (optimal growth temperature at 98 °C) shows an amino acid difference of 60% with respect to that of *C. symbiosum* (optimal growth temperature at 37 °C). It is obvious to expect that only a small fraction of such amino acid diversity is due to positive selection in response to high temperature. Another criticism concerns the assumption that all amino acid positions are equivalent and independent with respect to the substitution process, thus underestimating the importance of the local environment of a residue and the effects of amino acid substitutions at neighboring sites.

A better understanding of this process could be achieved by using predictive methods. As demonstrated by a series of recent publications (Chen et al., 2012a, 2013; Min et al., 2013; Xu et al., 2013a, 2013b; Xiao et al., 2013) and summarized in a comprehensive review (Chou, 2011), to develop an useful prediction method we need to consider the following procedures: i) construct a valid benchmark dataset to train and test the predictor; ii) consider a number of properties correlated with the attribute to be predicted; iii) introduce a powerful algorithm to operate the prediction; iv) perform a cross-validation test to evaluate the accuracy of the prediction.

Accordingly, we first constructed a training set of 33 thermo- and 36 mesophilic GDHs from Archaea and Bacteria. We then selected a number of physico-chemical properties of amino acids, some of them correlated to protein stability. We introduced the linear discriminant function (Fisher, 1936; Lachenbruch and Goldstein, 1979) as predictor of the determinants of thermostability. This choice was based on the ability of linear discriminant analysis to distinguish adaptive from neutral amino acid changes in protein evolution (Horimoto et al., 1990; Otsuka et al., 1993). Finally, we evaluated the accuracy of the prediction by using the Kolmogorov-Smirnov permutation test. This approach led to detection of 3 amino acid clusters as the putative determinants of thermal adaptation in GDH.

2. Materials and methods

2.1. Sequence data

We selected from the UniProt Knowledgebase (http://www.ebi.ac.uk/uniprot/) a training set of 69 GDH sequences. As shown in Table 1, it was subdivided into four groups: I) 17 GDHs from thermophilic Archaea with an optimal growth temperature (OGT) ranging from 55° to 98 °C II) 17 GDHs from mesophilic Archaea (OGT from 22° to 37 °C) III) 16 GDHs from thermophilic Bacteria (OGT from 55° to 80 °C) IV) 19 GDHs from mesophilic Bacteria (OGT from 27° to 37 °C). The training set included 7 GDHs encoded by paralogous genes: 2 from *Halobacterium salinarium*, 2 from *Haloarcula marismortui*, 1 from *Halogeometricum borinquense*, *Natronomas pharaonis*, and *Thermanaerovibrio acidaminovorans*, respectively. Information about OGT was taken from the German Collection of Microorganisms and Cell Culture (http://www.dsmz.de).

Table 1
List of the 69 GDH sequences of the training set.

Species	Ac. number	OGT (deg)	Length (aa)
Thermophilic Archaea (17 GDHs)			
Aeropyrum pernix	Q9YC65	90	418
Metallosphaera sedula	A4YIG2	65	421
Picrophilus torridus	Q6KZF2	55	415
Pyrobaculum aerophilum	Q8ZW33	98	427
Pyrobaculum calidifontis	A3MWK6	90	421
Pyrobaculum islandicum	A1RT74	97	428
Pyrococcus furiosus	P80319	98	420
Staphylothermus marinus	A3DLU5	88	426
Sulfolobus acidocaldarius	Q4JCA3	70	423
Sulfolobus solfataricus	Q97WS2	78	420
Sulfolobus tokodaii	Q96YC6	75	422
Thermococcus litoralis	Q56304	83	419
Thermococcus profundus	O74024	80	419
Thermococcus waiotapuensis	Q977 × 9	85	419
Thermofilum pendens	A1RYG4	88	419
Thermoplasma volcanium	Q97AN9	60	416
Thermoproteus neutrophilus	B1Y8Z9	85	427
Mesophilic Archaea (17 GDHs)			
Haloarcula marismortui	Q5V3Y8	37	418
Haloarcula marismortui	Q5V4 × 6	37	427
Haloarcula marismortui	Q5V6I7	37	431
Halobacterium salinarium	B0R2S8	37	429
Halobacterium salinarium	B0R537	37	416
Halobacterium salinarium	B0R3T5	37	417
Halobacterium halobium	P29051	37	435
Haloferax mediterranei	B9WPP9	37	417
Halogeometricum borinquense	E4NMU1	37	418
Halogeometricum borinquense	E4NMU2	37	431
Haloquadratum walsbyi	G0LKH3	37	419
Natronomas pharaonis	Q3IS94	37	424
Natronomas pharaonis	Q3ISK4	37	419
Methanocellula paludicola	D1YWP3	37	411
Methanococcoides burtonii	Q12UM4	23	416
Methanosarcina barkeri	Q467V9	35	419
Uncultured methanogenic archaeon RC-I	Q0W8B2	30	417
Thermophilic Bacteria (16 GDHs)			
Coprothermobacter proteolyticus	B5Y6Y1	60	416
Geobacillus thermodenitrificans	A4IQB6	60	423
Sphaerobacter thermophilus	D1C4B5	55	425
Sulfurihydrogenibium azorense	C1DWD7	68	418
Symbiobacterium thermophilum	Q67Q62	60	417
Thermoanaerobacter sp.	B0K183	67	416
Termoanaerobacter tengcongensis	Q8RAK8	75	413
Thermanaerovibrio acidaminovorans	D1B825	55	414
Thermanaerovibrio acidaminovorans	D1B8F3	55	424
Thermobaculum terrenum	D1CFL6	67	419
Thermomicrobium roseum	B9KZK6	70	421
Thermosinus carboxydivorans	A1HSF6	60	412
Thermosipho africanus	B7IG24	75	427
Thermosipho melanesiensis	A6LKL9	70	412
Thermotoga maritima	P96110	80	416
Thermus thermophilus	Q8GR86	75	419
Mesophilic Bacteria (19 GDHs)			
Acinetobacter baumanii	B0VQ84	30	423
Bacillus cereus	Q63DP1	30	428
Bacillus pumilus	A8FEN2	30	424
Bacillus subtilis	P39633	30	424
Bordetella pertussis	Q7VXC5	37	429
Collimonas fungivorans	Q6J667	27	428
Exiguobacterium sibiricum	B1YI38	28	421
Frankia sp.	A8LCL1	28	418
Klebsiella pneumoniae	A6T8K3	37	424
Lysinibacillus sphaericus	B1HXF2	30	414
Microscilla marina	A1ZDU1	22	424
Oceanobacillus iheyensis	Q8EQ98	28	426
Pelobacter propionicus	A1APQ5	30	420
Rhodococcus sp.	Q0RY06	28	423
Salmonella newport	B4SUI7	34	424
Staphylococcus aureus	Q6GID0	37	414
Serratia proteamaculans	A8GEN4	30	424
Sodalis glossinidius	Q2NWM5	28	423
Sporosarcina halophila	Q0E5I0	30	426