Protein Engineering vol.9 no.9 pp.725–726, 1996

FORUM

On the classification of modular proteins

Bernard Henrissat[1] and Peer Bork[2]

Centre de Recherches sur les Macromolécules Végétales[3], CNRS, BP53, F-38041 Grenoble, France and [2]EMBL, Meyerhofstr. 1, 69012 Heidelberg and Max Delbrück Center for Molecular Medicine, Berlin-Buch, Germany

[1]To whom correspondence should be addressed

[3]Affiliated with the Joseph Fourier University, Grenoble, France

A paper published recently by Wang *et al.* (1996) reports on the complementary use of five methods for the classification of protein sequences into related families. The family assignments of only five out of 16 823 proteins were different from those documented in PROSITE (Bairoch *et al.*, 1996) and Wang *et al.* (1996) concluded that this was due to a misclassification in PROSITE. The small number of differences in both classifications certainly points to the quality of the semiautomatic approach by Wang *et al.* (1996), but also speaks for the careful annotation within PROSITE. In fact, when inspecting the five examples (Wang *et al.*, 1996; their Table V), we found that all annotated PROSITE classifications appear to be correct (Table I).

Four of the five examples are multidomain proteins or proteins with multiple signatures (Table I). However, not all PROSITE signatures were linked to the four proteins in SWISSPROT, owing to the divergence of the respective sequences from the underlying signature patterns. Details are given here on only one example, the two-domain cellulase GUN5_THEFU; its catalytic domain clearly belongs to glycosyl hydrolase family 5 while its cellulose-binding domain (CBD) is related to other CBDs of the bacterial type (Figure 1).

Thus, the approach of Wang *et al.* (1996) is indeed complementary in that it is able to identify more divergent members of a PROSITE family (although all of these similarities have already been published). Wang *et al.* (1996) state that they take care of multidomain proteins: If 'a sequence belongs to multiple groups, ... a classifier is said to classify the sequence correctly if it hits any one of these groups'. Thus, Wang *et al.* (1996) should also have found the (correct) PROSITE annotations and only one example should have appeared as 'misclassified' in their Table V.

This fifth example, a modification methylase (SWISSPROT code MTB1_BACAM), is a more complicated case: annotated in SWISSPROT as N4-cytosine-specific methylase, the sequence matches the N6-adenine specific PROSITE signature (N6) and not the related one for N4-cytosine-specific methylases (N4). Phylogenetic trees based on multiple alignments including selected members of both N6 and N4 families as well as detailed inspection of the signatures confirmed (data not shown) a clustering of MTB1_BACAM to the N6 family. We conclude that the sequence classification in PROSITE is correct and the confusion might be caused by phenomena that are hard to trace such as independent evolution or an inappropriate functional characterization.

Proteins with a modular architecture are difficult to classify as many of the modules are just not known yet, others are not annotated owing to divergence and one should also be aware of misleading annotation due to inconsistencies in the nomenclature of modules (Bork and Bairoch, 1995). Automatic methods for domain identification and classification are desirable, but current methods still need the complementary input of expert knowledge.

References

Bairoch,A., Hofmann,K. and Bucher,P. (1996) *Nucleic Acids Res.*, **24**, 189–196.
Bork,P. and Bairoch,A. (1995) *Trends Biochem. Sci.*, **20**, Suppl., March, C03.
Wang,J.T.L., Marr,T.G., Shasha,D., Shapiro,B.A., Chirn,G.-W. and Lee,T.Y. (1996) *Protein Engng*, **9**, 381–386.

Table I. Domain architecture of the examples of Table V in Wang *et al.* (1996)

SWISS-PROT ID	Domains	Position	Annotation[a]
COG7_HUMAN	Propeptide	18–94	SW
	Zinc metalloprotease (PS00142)	95–297	SW,PS
	Matrixins/cysteine-switch (PS00546)	95–297	SW, PS
	Hemopexin repeats (PS00024)	297–467	W
GLNA_METVO	Gln synthetase: N-terminal tetrapeptide (PS00180)	1–446	SW, PS
	Gln synthetase: ATP-binding site (PS00181)	1–446	SW, PS
	Gln synthetase I specific : adenylation site (PS00182)	1–446	W
GUN5_THEFU	Cellulose-binding domain, bacterial type (PS00606	37–144	W
	Catalytic (glycosyl hydrolase family 5) domain (PS00659)	145–466	SW, PS
MCAS_MYCBO[b]	β-Ketoacyl synthase (PS00606)	1–?	W, SW[c], PS[c]
	Acyl transferase (active site: position 623)	?–?	SW
	Enoyl reductase (NADP site: position 1561)	?–?	SW
	β-Ketoacyl reductase (NADPsite: position 1765)	?–?	SW
	Acyl carrier (P-pantetheine attachment site; PS00012)	?–2100	SW, PS

[a]SW, annotated in SWISSPROT; PS, links from SWISSPROT to existing PROSITE entries; W, classified by Wang *et al.* (1996).
[b]The domain borders have not yet been determined as only functional sites are conserved. It is likely that even more domains will be detected in this multifunctional enzyme.
[c]Annotation added in November 1995.

© Oxford University Press

(a)

```
GUN5_THEFU  GPGTPSPDPGTQP-GTGTPVERYGKVQVCGTQLCDEHGNPVQLRGMSTHG  195
GUNA_STRLI  SVTDPPTDPPTDPPATGTPAAVNGQLHVCGVHLCNQYDRPIQLRGMSTHG  183
GUN1_BACSU  ----------PASAAGTKTPVAKNGQLSIKGTQLVNRDGKAVQLKGISSHG   66
GUNZ_ERWCH  ----------SSNAWASVEPLSVNGNKIYAGEKAKSFAGNSLFWSNNGWGG   78
                      | | || |* | *|    * |  |  ||||  || |  *

GUN5_THEFU  IQWFDHCLTDSSLDALAYDWKADIIRLSMYIQE-DGYETNPRGFTDRIDQ  244
GUNA_STRLI  IQWFGPCYGDASLDRLAQDWKSDLLRVAMYVQE-DGYETDPAGFTSRVNG  232
GUN1_BACSU  LQWYGDFVNKDSLKWLRDDWGITVFRAAMYTAD-GGYIDNPSV-KNKVKE  114
GUNZ_ERWCH  EKFYTA----DTVASLKKDWKSSIVRAAMGVQESGGYLQDPAGNKAKVER  124
             ||||     |||  *  **  || * |* ||| |**  |*   |||||

GUN5_THEFU  LIDMATARGLYVIVDWHILTPGDPHYNLDRAKTFFAEIAQRHASKTNVLY  294
GUNA_STRLI  LVDMAEDRGMYAVIDFHTLTPGDPNYNLDRARTFFSSVAARNDKK-NVIY  281
GUN1_BACSU  AVEAAKELGIYVIIDWHILNDGNPNQNKEKAKEFFKEMSSLYGNTPNVIY  164
GUNZ_ERWCH  VVDAAIANDMYAIIGWH---SHSAENNRSEAIRFFQEMARKYGNKPNVIY  171
             || * | ||*|||||*     ||| * | *  ** |||   ||| **|*

GUN5_THEFU  EIANEPNG-VSWA-SIKSYAEEVIPVIRQRDPDSVIIVGTRGWSSLGVSE  342
GUNA_STRLI  EIANEPNG-VSWT-AVKSYAEQVIPVIRAADPDAVVIVGTRGWSSLGVSD  329
GUN1_BACSU  EIANEPNGDVNWKRDIKPYAEEVISVIRKNDPDNIIIVGTGTWS------  208
GUNZ_ERWCH  EIYNEPL-QVSWSNTIKPYAEAVISAIRAIDPDNLIIVGTPSWS------  214
            ** ***   *|*  ||*|***|**||**  ***|||**** |**

GUN5_THEFU  GSGPAEIAANPVNASNIMYAFHFYAASHRDNYLNALREA-SELFPVFVTE  391
GUNA_STRLI  GANESEVVNNPVNATNIMYAFHFYAASHKDDYRAAVRPA-ATRLPLFVSE  378
GUN1_BACSU  -QDVNDAADDQLKDANVMYALHFYAGTHGQSLRDKANYALSKGAPIFVTE  257
GUNZ_ERWCH  -QNVDEASRDPINAKNIAYTLHFYAGTHGESLRNKARQALNNGIALFVTE  263
              | ||   ||||| *| *||****||* ||  |  | * ||  ||**|*

GUN5_THEFU  FGTETYTGDGANDFQMADRYIDLMAERKIGWT-KWNYSDDFRSGAVFQPG  440
GUNA_STRLI  FGTVSATAWSVDRSSSVAW-LDLLDQLKISYA-NWTYSDADEGSAAFRPG  426
GUN1_BACSU  WGTSDASGNGGVFLDQSREWLNYLDSKNISWV-NWNLSDKQESSSALKPG  306
GUNZ_ERWCH  WGTVNADGNGGVNQTETDAWVTFMRDNNIQLTQNWALNDKNEGASTYYPD  313
            |** | || |           || |   |*  | |*| |*   |||| *|
```

(b)

```
                signal peptide
            ----------------------------------
GUN5_THEFU  MAKSPAARKGXPPVAVAVTAA-LALLIALLSPGVAQAAGLTATVTKESSW   49
GUNC_PSEFL  MGHVTSPSKRYPASF-KRAGSILGVSIALAAFSNVAAAGC--EYVVTNSW   47
XYNB_PSEFL  MT-ISASDYRHPGNFLKRTTALLCVGTALTALAFNASAAC--TYTIDSEW   47
GUNA_CELFI  M-----STRRTAAALLAAAAVAVGGLTALTTTAAQAAPGCRVDYAVTNQW   45
            *     |   |       ||  |  |** | |  ||||   | | || *

GUN5_THEFU  DNGYSASVTVRNDTSSTVSQWEVVLTLPGGTTVAQVWNAQHTSSGNSHTF   99
GUNC_PSEFL  GSGFTAAIRITNSTSSVINGWNVSWQYNS-NRVTNLWNPNLSGSN-PYSA   95
XYNB_PSEFL  STGFTANITLKNDTGAAINNWNVNWQYSS-NRMTSGWNANFSGTN-PYNA   95
GUNA_CELFI  PGGFGANVTITN-LGDPVSSWKLDWTYTAGQRIQQLWNGTASTNGGQVSV   94
             |*||*|| | *  || || *||     |   |   **   ||||   |

GUN5_THEFU  TGVSWNSTIPPGGTASSGFIA---SGSGEPTHCTINGAPCDEGSEPGGPG   146
GUNC_PSEFL  SNLSWNGTIQPGQTVEFGFQGVTNSGTVESPT--VNGAACTGGTSSSVSS   143
XYNB_PSEFL  TNMSWNGSIAPGQSISFGLQGEKNGSTAERPT--VTGAACNSATTSSVAS   143
GUNA_CELFI  TSLPWNGSIPTGGTASFGFNGSWAGSNPTPASFSLNGTTCTGTVPTTSPT   144
            ||||**||*||* | | *| |   ||| |||   ||*||*|||  || ||
```

Fig. 1. (a) Multiple alignment of several glycosyl hydrolase family 5 sequences with the region corresponding to the catalytic domain of GUN5_THEFU; (b) multiple alignment of selected cellulose-binding domains (bacterial type) with the corresponding region of GUN5_THEFU. SWISS-PROT accession numbers for the various proteins in the alignment: GUN5_THEFU, Q01786; GUNA_STRLI, P27035; GUN1_BACSU, P07983; GUNZ_ERWCH, P07103; GUNC_PSEFL, P27033; XYNB_PSEFL, P23030; GUNA_CELFI, P07984. Identical residues are denoted with * and conservative replacements with |.