Journal of Bioinformatics and Computational Biology
Vol. 9, No. 1 (2011) 179–206

DOI: 10.1142/S0219720011005264

Imperial College Press
www.icpress.co.uk

THE JANUS-FACED E-VALUES OF HMMER2: EXTREME VALUE DISTRIBUTION OR LOGISTIC FUNCTION?

WING-CHEONG WONG[*,¶], SEBASTIAN MAURER-STROH[*,†,‖] and FRANK EISENHABER[*,‡,§,**]

[*]*Bioinformatics Institute (BII)*
*Agency for Science, Technology and Research (A*STAR)*
30 Biopolis Street, #07-01, Matrix, Singapore 138671

[†]*School of Biological Sciences (SBS)*
Nanyang Technological University (NTU)
60 Nanyang Drive, Singapore 63755

[‡]*Department of Biological Sciences (DBS)*
National University of Singapore (NUS)
8 Medical Drive, Singapore 117597

[§]*School of Computer Engineering (SCE)*
Nanyang Technological University (NTU)
50 Nanyang Drive, Singapore 637553
[¶]*wongwc@bii.a-star.edu.sg*
[‖]*sebastianms@bii.a-star.edu.sg*
[**]*franke@bii.a-star.edu.sg*

Received 16 July 2010
Revised 11 October 2010
Accepted 11 October 2010

E-value guided extrapolation of protein domain annotation from libraries such as Pfam with the HMMER suite is indispensable for hypothesizing about the function of experimentally uncharacterized protein sequences. Since the recent release of HMMER3 does not supersede all functions of HMMER2, the latter will remain relevant for ongoing research as well as for the evaluation of annotations that reside in databases and in the literature. In HMMER2, the E-value is computed from the score via a logistic function or via a domain model-specific extreme value distribution (EVD); the lower of the two is returned as E-value for the domain hit in the query sequence. We find that, for thousands of domain models, this treatment results in switching from the EVD to the statistical model with the logistic function when scores grow (for Pfam release 23, 99% in the global mode and 75% in the fragment mode). If the score corresponding to the breakpoint results in an E-value above a user-defined threshold (e.g. 0.1), a critical score region with conflicting E-values from the logistic function (below the threshold) and from EVD (above the threshold) does exist. Thus, this switch will affect E-value guided annotation decisions in an automated mode. To emphasize, switching in the fragment mode is of no practical relevance since it occurs only at E-values far below 0.1. Unfortunately, a critical score region does exist for 185 domain models in the hmmpfam and 1,748 domain models in the hmmsearch global-search mode. For 145 out the respective 185 models, the critical score region is indeed populated by actual sequences. In total, 24.4% of their hits have a logistic function-derived E-value

< 0.1 when the EVD provides an E-value > 0.1. We provide examples of false annotations and critically discuss the appropriateness of a logistic function as alternative to the EVD.

Keywords: Sequence homology; E-value; extreme-value distribution; logistic function; HMMER2; Pfam; sequence annotation.

1. Introduction

Sequencing of DNA has become the key life science research technology only because computational methods provide an opportunity for the functional characterization of otherwise not (especially not experimentally) studied genes and protein molecules. The transfer of functional annotation from an experimentally characterized example to a whole family of proteins with similar sequences is justified by the theory of sequence homology. Assuming a common ancestor and evolutionary divergence due to mutational events, selection pressure for biological function will, as a trend, result in similarity of amino acid sequence, three-dimensional structure and molecular function for all members of the protein family.[1–4]

To enhance the sensitivity in sequence similarity searches, it is necessary to apply sophisticated profile searches[5,6] embedded in complex search heuristics.[7,8] Since protein segment family collection, their alignment and the subsequent profile generation represent a considerable effort, domain libraries have become an indispensable tool for annotation of uncharacterized sequences. Among the publicly available collections, most notable are BLOCKS,[9] CDD,[10] EVEREST,[11] libraries associated with IMPALA,[12] PANTHER,[13] PRINTS,[14] ProDom,[15] PROSITE,[16] SUPERFAMILY[17] and, as the most used primary ones, Pfam[18,19] and SMART.[20]

Many domain libraries provide protein domain models in the form of hidden Markov models (HMMs).[6,21] There is now more than a decade of empirical experience of using the program HMMER2[6,21] for similarity searches with models mainly from Pfam and SMART. This technology is tremendously helpful and has become the cornerstone for annotating fully sequenced genomes. It should be noted that the recent release of HMMER3[22,23] does not override HMMER2. HMMER3 only partially substitutes for HMMER2 since (i) it has only a fragmentary but no global domain search variant.[22,24] (ii) HMMER3 is not tested to the extent of HMMER2 with regard to accuracy whereas the application of the latter has a record of a decade of important biological discoveries. Importantly, the hit lists of HMMER2 and HMMER3 are overlapping yet not identical. It is not clear at present whether HMMER3 in its present form will really become the mainstream in domain prediction. (iii) So far, only Pfam has changed to HMMER3 but not other domain databases such as SMART. Further, domain assignments in many sequence databases and in the literature have been generated with HMMER2 and have not been and will not be renewed with HMMER3. Thus, HMMER2 has some more time to live and understanding its way of E-value assignment remains relevant. The issue of exaggerated E-values by HMMER2 has been noticed empirically

by many in the community[22,25]; yet, the reasons remained unclear. Although late, this article resolves part of the mystery.

Correctness of a hit of an HMM model within a protein query sequence can either be taken with the gathering score criterion or be E-value guided.[6,21] The gathering score, the lowest score of a known good hit without false positives having higher scores (at the time of model construction), is a conservative criterion; yet, it misses many good hits. We have extensively discussed the deficiencies of the gathering score approach elsewhere.[26] E-value guided annotation transfer allows deeper extrapolation into the sequence space when, at the same time, the false-positive error remains statistically evaluated. In the manual provided with HMMER2, Sean Eddy advises that "The best criterion of statistical significance is the E-value. The E-value is calculated from the bit score. It tells you how many false positives you would have expected to see at or above this bit score. Therefore a low E-value is best; an E-value of 0.1, for instance, means that there's only a 10% chance that you would've seen a hit this good in a search of non-homologous sequences. *Typically, I trust the results of HMMER searches at about* $E = 0.1$ *and below, and I examine the hits manually down to* $E = 10$ *or so.*"

When using HMMER2-style HMMs, we anecdotally observed the trend for extremely low E-values for known good hits; yet, very large E-values (maybe 50 orders of magnitude higher) for sequences that still share some stretches of similarity with the model and little sampling of the E-value space in between (e.g. see Supplementary File 1; supplementary files for this article are also available via http://mendel.bii.a-star.edu.sg/SEQUENCES/ProblemDomains-JanusEvalue). In contrast to the vast changes in E-values, the respective scores are not so different. Apparently, a number of factors appear responsible for this behavior. For example, the non-redundant database provides limited sequence sampling, the parameters of the EVD are not well estimated and this is aggravated in cases of long domains.

There is also a technical reason that we wish to analyze in this article: For our previous work,[26] we needed to compute sequence segment-based contributions to the total HMM-derived score. As a control, we tried to reproduce E-values generated with HMMER2 over a wide range of conditions. Surprisingly, we found a systematic divergence in E-values for large scores (Fig. 1). When checking the HMMER2 code, we stumbled onto the switching between two statistical models for E-value generation (routine *P*-Value in "Mathsupport.c", see Supplementary File 2, comments in red). In essence, two *p*-values are concurrently calculated in this piece of code; one from the extreme-value distribution (EVD) and another from the logistic function. However, the smaller *p*-value is always selected for the final computation of the E-value (*p*-value multiplied by the size of the sequence/domain database). Thus, for some scores, E-values are calculated with an EVD; for other scores beyond a certain log odd score $S_{\text{breakpoint}}$ (see methods for its computation), a logistic function is applied (Fig. 1). Since both functions used for E-value computation are monotonous, the breakpoint represents a point of switching between statistical models.

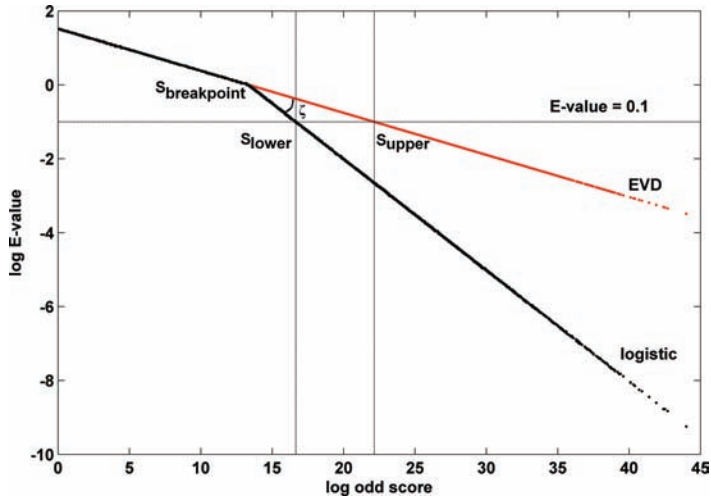

Fig. 1. Divergence plot of E-values with respect to the EVD (extreme value distribution) and logistic function beyond the breakpoint score $S_{\text{breakpoint}}$. For majority of the Pfam release 23 domain models in global-mode hmmpfam/hmmsearch mode, the logistic function will supercede the EVD for their E-value calculations beyond a breakpoint [see Eq. (8)] along the positive score axis. Furthermore, some of these domains contain a critical region bounded by the lower and upper score [Eqs. (13) and (14)] where hits in these regions are contradicted by the two statistical models.

The inclusion of the logistic function into the E-value computation raises several issues for practical (automated) sequence annotation and theoretical justification. From a practical perspective, there will be a range of scores [the interval $(S_{\text{lower}}, S_{\text{upper}})$, Fig. 1] for a number of domain models where some hits look insignificant based on EVD (given, for example, the threshold of 0.1) but become significant in the present HMMER2 framework since the logistic function produces more exaggerated E-values. From a theoretical perspective, having a hybrid of two statistical models (EVD and logistic function) complicates statistical inference of sequence similarity for homology. The comparability of E-values for hits matching the same sequence region becomes problematic since the E-values might be generated under two different statistical models. Furthermore, the logistic function as a statistical model to measure sequence similarity has not been justified fundamentally. The subsequent sections will address these issues in greater detail.

2. Results

2.1. *The overwhelming majority of domains in Pfam (release 23) switch between EVD- and logistic function-derived E-values in HMMER2 searches depending on score value*

We computed the critical score $S_{\text{breakpoint}}$ for the Pfam domain models in release 23, the last version dedicated to HMMER2 [see Sec. 4, especially Eqs. (8) and (10)]. For

this purpose, the respective EVD parameters λ and μ [which are in fact search-mode dependent and are different for the ls- (global) and fs- (fragment) domain searches] were extracted from the Pfam HMMs.

We found that, for the ls- (global-) mode, an overwhelming majority (10,337 out of 10,340) of domain models in Pfam has a positive breakpoint [and $\lambda \ll \log 2$, see Eq. (8)]. In these cases, the EVD is used for E-value computation for scores below the breakpoint and is substituted by the logistic function for larger scores, a function with considerably steeper deceleration in E-value (Fig. 2).

The only exceptions with a negative breakpoint solution are PF02095.7 (Extensin-like_protein_repeat), PF06049.4 (Coagulation_Factor_V_LSPD_Repeat) and PF07391.3 (NPR_nonapeptide_repeat) (their $\lambda > \log 2$; see Fig. 2). These three cases are special since the logistic function always delivers a P-value that is larger or equal to the P-value from the EVD. Thus, the EVD is the only statistical model used for E-value calculation and the issue of switching statistical models is not relevant for these three models.

For the fs-mode, there are 2,136 domains with negative breakpoint ($\lambda > \log 2$) and another 415 domains with a $\lambda \approx \log 2$ (between 0.6849 and 0.6931) resulting in

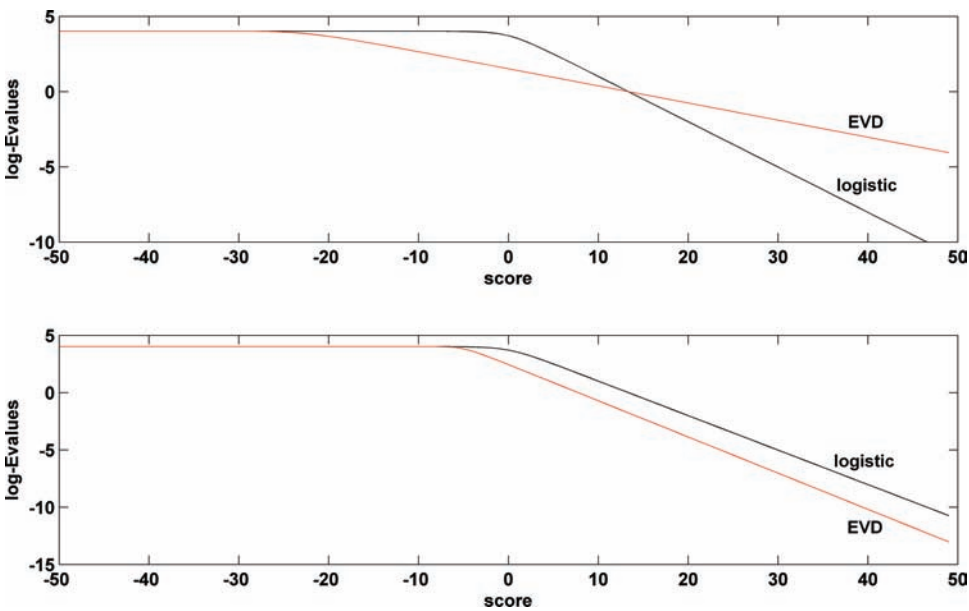

Fig. 2. Plots of EVD (extreme value distribution) and logistic function for a domain with parameter $\lambda < \log 2$ and $\lambda > \log 2$ respectively. Based on the upper figure, for domains with EVD parameter $\lambda < \log 2$, the EVD (in red) and logistic function (in black) will cross over at some point on the positive score axis. We defined this as the breakpoint. When this happens, the E-value calculations will be based on the logistic function which is more aggressive in nature. On the other hand, for domains with EVD parameter $\lambda > \log 2$, cross-over between the EVD and logistic will not occur. Thus for these models, the E-value calculations are dependent only on the EVD (see lower figure).

a breakpoint with very large positive score value. In both cases, the EVD is below the logistic function over the full range and the only statistical function that is used for calculating E-values. For the remaining 7,789 domain models ($\lambda < \log 2$), all switch from the EVD to the logistic function as score increases beyond a positive breakpoint.

2.2. *Hundreds of domains in Pfam (out of 10,340 in release 23) have a score interval with conflicting EVD/logistic function-derived E-values based on global-mode HMMER2 searches*

The switching between statistics for E-value calculation remains non-critical from the annotation point of view if the E-value corresponding to $S_{\text{breakpoint}}$ is lower than the E-value threshold used for deciding true domain model hits (here and throughout this work, the threshold is 0.1, the value recommended by Sean Eddy in the manual provided with HMMER2). In cases where the breakpoint generates an E-value that is larger than the established threshold, a critical region $(S_{\text{lower}}, S_{\text{upper}})$ [Eqs. (13) and (14)] exists. More specifically, if a hit is being flagged significant by the EVD, it is also significant based on the logistic function since the latter produces a more extreme E-value with the caveat that the error measure (i.e. E-value) becomes more impressive (Fig. 1). Meanwhile, the converse argument is not true. Annotation decisions based on the logistic function may result in an underestimation of false-positive hits since more hits become insignificant when evaluated by the EVD. This issue is relevant for automated E-value guided extrapolations in annotation pipelines.

Figure 3 depicts the histogram of breakpoint E-values (in logarithmic scale) for all domain models in Pfam release 23 based on global-mode searches. The result is influenced by the size of the database [Eq. (11) in Sec. 4] and the cases of "hmmpfam" (with database size equal to the number of models) and "hmmsearch" (with database size equal to the number of sequences in the non-redundant database) should be distinguished. To recall, both EVD and the logistic function give the same E-value at the breakpoint.

In the "hmmpfam" mode, we find that the median E-value in Fig. 2 is in the order of 1.e-7. In total, 185 models have a critical score $S_{\text{breakpoint}}$ corresponding to an E-value larger than 0.1 (Fig. 3 and see Supplementary File 3 for their list) and, hence, these models give rise to the critical region $(S_{\text{lower}}, S_{\text{upper}})$. When tested with all query sequences from the non-redundant database (downloaded on 5th April 2010, n = 10,818,955), we found that, out of the 185 models, this region is sampled by actual sequences in 145 models and, on average, 24.4% of all hits with E-value below 0.1 (measured by the logistic function) belong to that interval. In the "hmmsearch" mode, we used the same non-redundant database and found 1,748 domain models with $S_{\text{breakpoint}}$ corresponding to an E-value above 0.1 (Fig. 3 and as a list in Supplementary File 3). Not surprisingly, the 185 domains from the

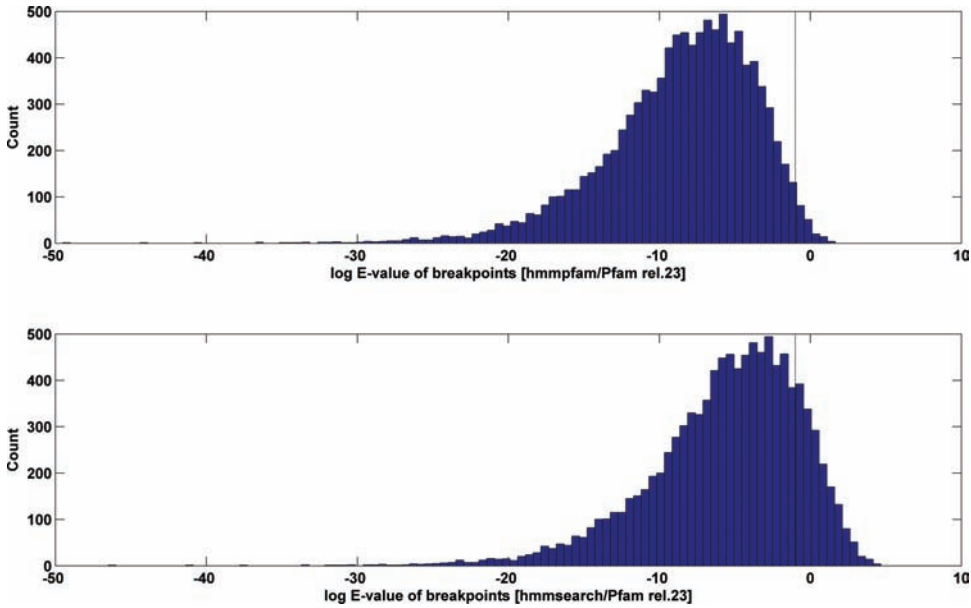

Fig. 3. Histograms of breakpoint E-values from global-mode hmmpfam/hmmsearch on Pfam release 23. The figure above depicts the histogram of log E-values of breakpoints for global-mode hmmpfam of Pfam release 23. In this case, the database size is 10,340. At an E-value of above 0.1, 185 Pfam models contain the critical region where the logistic function suggests the hits as being significant, yet the EVD says otherwise. Similarly, the figure below depicts the histogram of the log E-values of breakpoints for global-mode hmmsearch where the database size is 10,818,955. Given an increase in database size, more Pfam models are expected to contain the critical region even at the same E-value cutoff. In this case, this number increased to 1,748.

"hmmpfam" case form a subset of the 1,748 domains in the "hmmsearch" case. The principal difference between the two search modes is the size n of the database [see Eqs. (11)–(14) in Sec. 4]. As the database size increases over time, more domains are expected to acquire a critical region. The number of domains with a critical region depends on the database size. Thus, in these cases, an automated, E-value guided decision making procedure would be affected by the change of the statistics for error estimation.

Alternatively, a so-called gathering score, an expert-defined score threshold with the lowest score of a known good hit without known false-positives with higher score (at the time of model creation) is advised to decide between good hits and false matches. It should be noted that, for 37 out of the 185 domain models (in the "hmmpfam" mode) and 66 out of 1,748 domain models (in the "hmmsearch" mode), the gathering score is within the respective critical regions. Also, for 19 out of 185 (in the "hmmpfam" mode), their gathering scores are even less than S_{lower}. In the "hmmsearch" mode, this is the case for 1,651 domain models out of the respective 1,748. In the case of automated annotation assignment based on the

gathering score approach, the accompanying E-value would be computed via the EVD up to the breakpoint and via the logistic function for scores above $S_{\text{breakpoint}}$. Thus, using gathering scores does not ensure calculation of E-values with one and the same statistical model.

In contrast to the global-mode searches, switching between statistical models is not an issue for the fragment-mode searches. As a trend, the parameter λ for fragment mode is larger than that of the global mode. This gives rise to larger breakpoints that, in turn generate smaller P- and E-values. Though switching between EVD and logistic function still exists for nearly 80% of all domain models, the larger breakpoints shift the switch into a region where both EVD and logistic function dive below the E-value threshold of 0.1. Thus, the issue is of no practical relevance here, although, of course, the logistic function generates considerably smaller E-values than the EVD in this region.

2.3. *Some examples of likely false-positive hits with only logistic function-derived E-value support in the hmmpfam global-mode search*

Since the logistic function-derived E-values more optimistically evaluate the significance of a domain model hit, we especially searched for sequence examples that are supported by the E-value derived from the logistic function but not from the one calculated with the EVD approach. We wanted to know whether this discrepancy would lead to likely annotation errors only supported by the logistic function-derived E-value in an automated mode. Indeed, such examples do exist.

For the beginning, we focused our search on domain models with 3D structural support. To further reduce the number of domain model hits for manual screening, we first also required that the score of such sequence examples was below or close to the respective gathering score of the domain model, hereby following the assumption that all larger-score hits are likely correct. At the stage of manual handling, we evaluated the alignment quality (especially, of the hydrophobic pattern) and the taxonomic diversity of the sequences with the potential hit (domain architectures that occur just for a single sequence more likely indicate false matching). Further, function contradiction of the apparently false-positive hit overlapping with a more significant domain hit provided further evidence for the conclusion. Exemplary results from this search are provided in Table 1 (listed with ascending entry number) and Fig. 4 (the respective alignments are shown in Supplementary File 4).

Sequence XP_001373560.1 appears to be a sugar transferase; yet, the cytochrome B559 domain (PF00283.11, from photosystem II) hits into the sequence providing a function conflict. The next sequence (ZP_01450808.1) contains a seryl-tRNA synthetase domain but this annotation is conflicted by another domain (PF00627.23 UBA, an ubiquitin associated domain) at a lower significance. Similarly, the ABC transporter domain of sequence YP_796932.1 is also in conflict with a domain

Table 1. Summary of likely false-positive hits in selected Pfam domains (detected by global-mode hmmpfam runs). The first column provides the Pfam entry number (release 23) and an excerpt from the domain model description. In the second column, a representative structure for the given model is listed together with the literature reference. The accessions of protein sequences (together with a running example number and the sequence length) as well as comments from their database description are presented in columns three and four. The last column provides the following pieces of information: (i) the score, (ii) the E-value of the Pfam mode hit returned by hmmpfam (HMMER2; i.e. the E-value from the logistic function), (iii) the number of domain hits, and (iv) the E-value for the same hit from the EVD.

Domain name	Reference structure PDB id	Sequence accession no. (no. of AA)	Sequence description/taxonomy	Hmmpfam score/ E-value/No.of hits (EVD E-value)
PF00283.11: Cytochrome_B559 (Cytochrome b559, alpha (gene psbE) and beta (gene psbF)subunits) GA: 25; alignment length: 29; HMM length: 29	1IZL_F Ref. 44	1. XP_001373560.1 (479AA)	PREDICTED: similar to beta-1,4-N-acetyl-galactosaminyl transferase 2, *Monodelphis domestica*	16.9/8.4e-2/1 (1.3e-1)
PF00627.23: UBS (UBA/TS-N domain) GA: 22.4; alignment length: 46; HMM length: 38	1OQY_A Ref. 45	2. ZP_01450808.1 (88AA)	Seryl-tRNA synthetase, *alpha proteobacterium HTCC2255*	18.0/4e-2/1 (1.1e-1)
PF01402.13: RHH_1 (Ribbon-helix-helix protein, copG family) GA: 25.2; alignment length: 40; HMM length: 40	2CPG_A Ref. 46	3. YP_796932.1 (629AA)	ABC transporter permease/ATP-binding protein, *Leptospira borgpetersenii serovar Hardjo-bovis L550*	16.8/9.1e-2/1 (2.0e-1)
PF01814.5: Hemerythrin (Hemerythrin HHE cation binding domain) GA: 23.5; alignment length: 110; HMM length: 70	2AVK_A Ref. 47	4. XP_391082.1 (302AA) 5. XP_781670.1 (708AA)	Hypothetical protein FG10906.1, *Gibberella zeae PH-1* PREDICTED: similar to adenylate kinase 7, *Strongylocentrotus purpuratus*	17.6/5.4e-2/1 (3.3e-1) 17/8.0e-2/1 (3.9e-1)
PF01842.17: ACT (ACT domain) GA: 18.6; alignment length: 81; HMM length: 64	2DTJ_A Ref. 48	6. YP_233518.1 (540AA)	Hypothetical protein Psyr_0410, *Pseudomonas syringae pv.syringae B728a*	18.1/3.8e-2/1 (1.1e-1)
PF01844.15: HNN (HNN endonuclease) GA: 22.5; alignment length: 133; HMM length: 58	1M08_A Ref. 49	7. YP_001633696.1 (549AA)	Methyltransferase type 11, *Chloroflexus aurantiacus J-10-fl*	17/7.7e-2/1 (1.6e-1)
PF04151.7: PPC (Bacterial pre-peptidase C-terminal domain) GA: 21.8; alignment length: 175; HMM length: 85	1WMD_A Ref. 50	8. YP_433983.1 (3258AA) 9. NP_521905.1 (221AA)	Fibronectin type III domain-containing protein, *Hahella chejuensis KCTC 2396* basal-body rod modification protein FlgD, *Ralstonia solanacearum GMI1000*	17.5/5.7e-2/1 (1.3e-1) 16.8/9.2e-2/1 (1.6e-1)

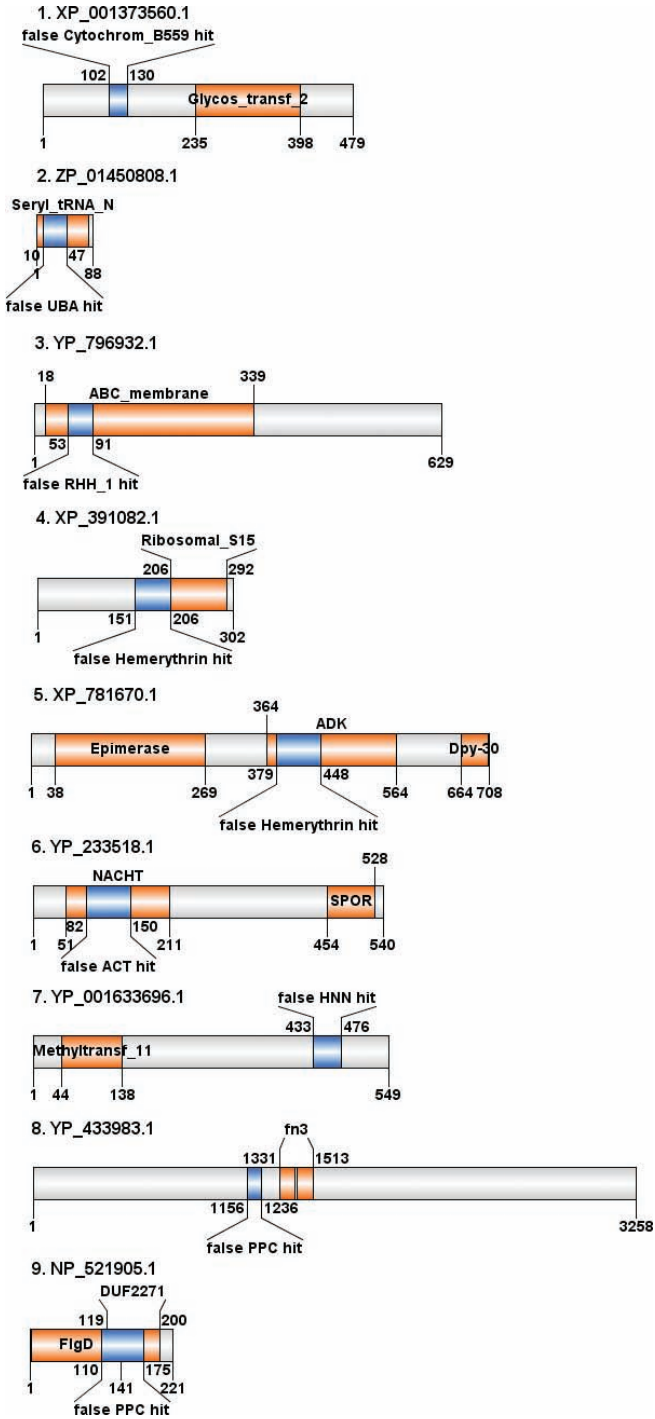

Fig. 4. Domain architecture of selected likely false-positive hits (detected in the hmmpfam mode). The illustrations refer to accession numbers and descriptions in Table 1.

hit (PF01402.13 RHH_1, a Ribbon-helix-helix protein of copG family) of lower significance.

The next two examples of an isolated hemerythrin domain (actually half of the domain is encoded in PF01814.5; thus, two subsequent hits are required) in XP_391082.1 (besides the S15 ribosomal domain) and in XP_781670.1 (hitting two structural helices of a kinase domain) are clearly also false positives since the other domain half cannot be placed into the sequence. Similarly, the YP_233518.1, a sporulation factor, has an isolated ACT (PF01842.17) hit when two subsequent ones are required for the full domain and this hit conflicts with a NACHT/PF05729.4 domain.

YP_001633696.1 is an example of a function conflict, a methyltransferase with a false-positive endonuclease domain (PF01844.5) hit. To note, the respective alignment has a large gap including the 3_{10}-helical segment in front of a β-strand. Finally, the bacterial pre-peptidase (C-terminal domain, PF04151.7) should be observed in concert with its N-terminal; yet, its isolated hit is a false-positive both in YP_433983.1 (a fibronectin) and in NP_521905.1 where it covers the boundary region of two other neighboring domains (the strong hits of FlgD/PF03963.6 and DUF2271/PF10029.1).

2.4. *More examples of likely false-positive hits with only logistic function-derived E-value support in the hmmsearch global-mode search*

The following exemplary false-positive sequences were first isolated by hmmsearch runs and then re-annotated via hmmpfam (see Table 2 and Fig. 5, alignments are provided in Supplementary File 5). It should be noted that all these examples produce formally impressive E-values since hmmsearch (and hmmpfam) routes the score into the logistic function routine; yet, these examples would escape any attention if only the EVD-based E-value calculation would be applied. For selection of the examples, we applied criteria similar to the ones in the previous section.

The first example, the sequence NP_001146906.1 is clearly an apurinic endonuclease-redox protein. Nevertheless, the domain SAP (PF02037.19) has a single, apparently significant but false-positive hit in this sequence; it should be noted that two of such hits would be required for a full SAP domain. Though its hmmpfam annotation includes an additional insignificant SAP domain (that overlaps with the true hit of the "Exo_endo_phos" domain model PF03372.15), its overall score/E-value is made significant artificially by the logistic function.

In the case of sequences YP_002512543.1 and ZP_05026152.1, both are serine/threonine protein kinases. They have a single, weak HEAT_PBS domain hit (PF03130.8, found in phycobilisomes (PBS) that are peripherally attached to the photosynthetic membrane). In the hmmpfam mode though, there is a large number of such hits that add up for an impressive E-value. Nevertheless, they are false-positives since they provide a function contradiction.

Table 2. Summary of likely false-positive examples in selected Pfam domains (detected by global-mode hmmsearch runs). The first column provides the Pfam entry number (release 23) and an excerpt from the domain model description. In the second column, a representative structure for the given model is listed together with the literature reference. The accessions of protein sequences (together with a running example number and the sequence length) as well as comments from their database description are presented in columns three and four. The fifth column provides the following pieces of information: (i) the score, (ii) the E-value of the Pfam mode hit returned by hmmsearch (HMMER2; i.e. the E-value from the logistic function), (iii) the number of domain hits, and (iv) the E-value for the same hit from the EVD. In the last column, the same four values are presented if the hits are computed with hmmpfam.

Domain name	Reference structure PDB id	Sequence accession no. (No. of AA)	Sequence description/ taxonomy	Hmmsearch score/ E-value/No.of hits (EVD E-value)	Hmmpfam score/ E-value/No.of hits (EVD E-value)
PF02037.19: SAP (SAP domain)	1JJR_A Ref. 51	1. NP_001146906.1 (510 AA)	Apurinic endonuclease-redox protein, *Zea mays*	35.8/1.7e-4/1 (1.15e-1)	38.6/2.5e-08/2 (5e-5)
GA: 28.8; alignment length: 35; HMM length: 35					
PF03130.8: HEAT_PBS (PBS lyase HEAT-like repeat)	1TE4_A Ref. 52	2. YP_002512543.1 (593 AA)	Serine/threonine protein kinase, *Thioalkalivibrio sp. HL-EbGR7*	26.7/1e-1/1 (1.2)	103.1/9.8e-28/7 (3.8e-14)
GA: 27.9; alignment length: 40; HMM length: 27		3. ZP_05026152.1 (1,184 AA)	Ser/Thr protein phosphatase family protein, *Microcoleus chthonoplastes PCC 7420*	29.4/1.5e-2/1 (4.7)	143.3/7.3e-40/8 (3.5e-20)
PF03958.9: Secretin_N (Bacterial type II/III secretion system short domain)	3EZJ_A Ref. 53	4. ZP_01993497.1 (356 AA)	4-diphosphocytidyl-2-C-methyl-D-erythritol kinase, *Vibrio parahaemolyticus AQ3810*	42.4/1.8e-06/1 (3.8e-1)	42.4/1.7e-09/1 (3.7e-4)
GA: 29.5; alignment length: 166; HMM length: 76					
PF07498.4: Rho_N (Rho termination factor, N-terminal domain)	1A8V_A Ref. 54	5. XP_001638172.1 (1617 AA)	Predicted protein, *Nematostella vectensis*	27/7.8e-2 /1 (6.8e-1)	27/7.5e-05/1 (6.5e-4)
GA: 31; alignment length: 51; HMM length: 43		6. ZP_03700647.1 (174 AA)	Ribosomal protein L21,*Flavobacteria bacterium MS024-3C*	26.9/8.8e-2/1 (7e-1)	26.9/8.4e-05/1 (6.7e-4)
PF07554.5: FIVAR (Uncharacterised Sugar-binding Domain)	2DGJ_A Ref. 55	7. ZP_03708976.1 (1480 AA)	Hypothetical protein CLOSTMETH_03737,*Clostridium methylpentosum DSM 5476*	34.6/4.2e-4/1 (2.1)	88.1/3.2e-23/3 (4.9e-7)
GA: 26.8; alignment length: 71; HMM length: 64					

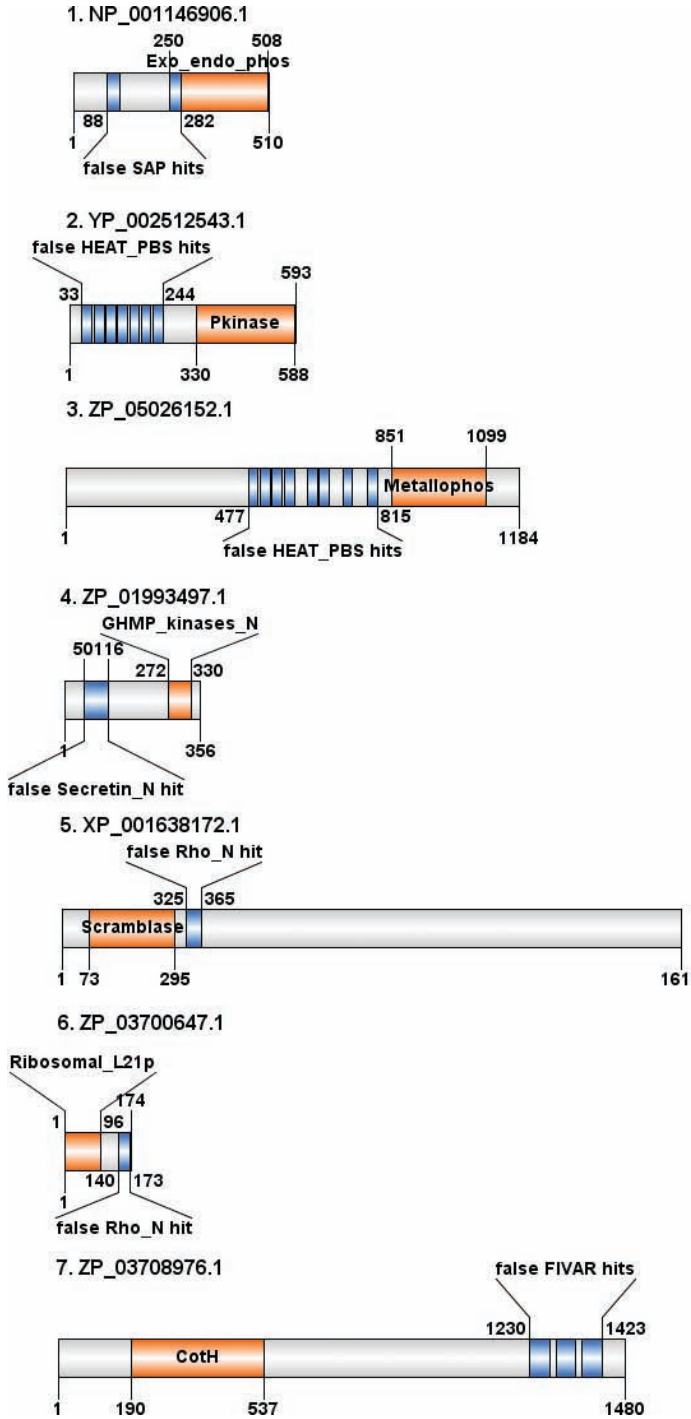

Fig. 5. Domain architecture of selected likely false-positive hits (detected in the hmmsearch mode). The illustrations refer to accession numbers and descriptions in Table 2.

Next, the 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase, ZP_01993497.1 has an apparently significant, nevertheless false Secretin_N hit (PF03958.9, the Secretin_N domain hits should occur in tandem and there is a function mismatch, too) which was expectedly more significant in the hmmpfam mode. Please note the large difference in E-value: 1.8e-6 for the logistic function and 0.38 for the EVD.

The next two sequences XP_001638172.1 and ZP_03700647.1 (a scramblase and a ribosomal protein) hit a Rho_N domain (PF07498.4) at their C-terminus (when the N-terminal should be expected) without the Rho_RNA_bind domain (PF07497.4) and, hence, suggest function contradiction.

Finally, ZP_03708976.1 is a spore coat protein H with a single FIVAR domain hit (PF07554.5, two FIVAR domains are required) and, again, its hmmpfam score/E-value is artificially exaggerated due to the inclusion of two insignificant FIVAR domain hits.

3. Discussion

3.1. *EVD as the correct statistic to evaluate the significance of sequence alignments*

The question of statistical significance of an alignment between two biomolecular sequence segments with reference to a user-defined scoring scheme (or of a sequence and a profile) has been the center of attention within the computational biology community for many decades.[27–29] There are two principal ways of approaching the problem — by generating databases of random sequences or by deriving closed form expressions for the asymptotes of the statistical distribution. The most remarkable result of this research is the finding that an extreme-value distribution is the statistic that characterizes the significance of local ungapped alignments.[30,31]

On this basis, BLAST[32] has become the first sequence similarity search tool with statistical significance estimation of hits via E-values and, for the first time, it became possible to objectively assess the validity of generated alignments. It should be noted that there is some dependence on the monomer composition in the sequences; the distribution changes for biased compositions what can be corrected for with some modifications.[33–37] From the practical point of view, suppression of sequence regions with biased composition or simple repeats before submission to sequence similarity searches is still the better alternative since it provides improved sensitivity. In this way, false-positive matches due to biased regions are *a priori* excluded.[8,26]

3.2. *Cause of divergence in HMMER2 E-values from the EVD distribution and the issues of E-value generation by the logistic function*

The HMMER package does make use of the improved scientific understanding of significance estimates for profile-sequence matches. The EVD is included as an

option for computing E-values. In addition, a simpler formulation with the logistic function (called "sigmoid" by Sean Eddy on page 48 of the User Guide for HMMER2 from October 2003) is used. In the comment lines of the HMMER2 source code (mathsupport.c, see Supplementary Fig. 1), the EVD is called the tighter bound. Thus, the creator of HMMER2 considered that the jumping between two statistical models has no real practical importance since, EVD-based E-values will be used as a rule. As we know now, this is indeed true for the fragment mode both in hmmsearch and hmmpfam. Even if there is a region where the logistic function-based E-values are smaller than those calculated with EVD (for about 75% of all models, such regions do exist at large positive breakpoints $S_{\text{breakpoint}}$), this happens only in regions with very large scores where both statistical models generate E-values below 0.1.

Surprisingly, this is different for the global mode. As a trend, the parameters λ of the EVD are smaller for the global mode than for the fragmented mode; thus, there are more cases of domain models for which switch points $S_{\text{breakpoint}}$ correspond to sufficiently low scores so that they can give rise to critical score regions [see Sec. 4, Eq. (12)]. In Tables 1 and 2 (see also Figs. 4 and 5), we provide some examples that, for the global mode, the switching between statistical models indeed influences annotation results. Of course, all false-positive hits of this kind can be suppressed by manual intervention. Yet, this switching from the EVD to the logistic function is a cause of systematic errors in an automated sequence annotation system.

To note, the sensitivity of the domain model (measured as steepness of decrease of E-values for growing scores) is dependent on the EVD parameter λ. For scores sufficiently large (where $e^{s \log 2} \gg 1$), the logistic function mimics an EVD with parameters $\mu = 0$ and $\lambda = \log 2$. Hence, for those Pfam models with the switching to the logistic function, their final λ after the switch are fixed at $\log 2$. In addition, since global models on average have smaller switch points $S_{\text{breakpoint}}$ than fragment ones (42 versus 197), all global models (except for PF02095.7, PF06049.4 and PF07391.3) adopt the forced fed $\lambda = \log 2$ relatively early on the score axis discarding their model-specific λ. Taken together, it is not surprising that the global Pfam models are reportedly more sensitive than the fragment mode models despite their gentler EVD parameters of $\lambda_{\text{global}} \sim N(0.191, 0.068)$ (normal distribution with mean and standard deviation in parentheses) as compared to $\lambda_{\text{fragment}} \sim N(0.668, 0.051)$.

3.3. *Validity of the logistic function in HMMER2 to evaluate the significance of sequence similarity*

Essentially, it appears to us that there is no clear reason why a logistic (or "sigmoid") function should be included into the E-value calculation routines other than for cosmetically increasing the sensitivity of the domain models. This is especially true for domain models that generate lower scoring hits in the global mode.

To emphasize, the domain model parameters generate, as a trend, lower scores in the global mode compared with the fragment mode (arithmetic averages of the μ parameters over all domain models in Pfam release 23 result in -149.1 and -10.0 for the global and fragment modes respectively whereas the corresponding medians are -110.9 and -10.0). Thus, there is no clear advantage visible from the practical point of view.

For high scores of hits that are anyway significant, the logistic function creates the impression of excessive significance by creating E-values of several orders of magnitude lower than that of the EVD. We have seen that, in the twilight regions for some domain models, the logistic function-based E-values support a confidence in hits that is unfounded. Maybe, it was thought in the first years of HMMER that EVD parameters would not be readily available for all domains and, in these cases, the logistic function could supply some rough E-value substitute.

And there is also no theoretical argument in support for the jumping between statistical models. As the distribution of a sum of independent, identically distributed random variables is normally distributed, the maxima of independent, identically distributed random variables are extreme-value distributed. Not surprisingly, optimized scores of matches between unrelated random sequences follow the EVD. The EVD has been demonstrated to reliably estimate significance of sequence matches both via theoretical derivation and numerical experiments.[30,31] There is no body of theory behind the logistic function for this purpose.

In contrast to EVD, the logistic function is symmetrical with regard to low and high scores. The rationale is straightforward; given the same score, the area under the curve for the right-hand side of the logistic function is smaller than that of the EVD since it does not extend as far into the infinity score axis, thus always resulting in a smaller P-value. Conversely, the area under the curve for the left-hand side of the EVD is smaller than that of the logistic function since the total area must add to one. As a result, the E-value computation in HMMER2 is essentially based on a hybrid p-value distribution composed of the left-hand side of an EVD and the right-hand side of a logistic function.

It is known that the difference of two maximum Gumbel distribution (Type I EVD) equates to the logistic function.[38] Thus, insisting on the relevance of the logistic function as statistic requires arguments for justification of a second EVD. This appears a problematic endeavor. Finally, using a hybrid distribution model (switching from EVD to logistic function with increasing score) without clear resolution of the fundamental problems questions the comparability of E-values computed for different Pfam models over the same sequence region. A statistician might consider the hybrid distribution a betrayal of pure tenets; yet, we wish to emphasize that the creator of HMMER has rightfully not seen this being a big issue since, indeed, the fragment mode is practically unaffected.

3.4. *About the domain model-specific divergence between the EVD and the logistic function*

With reference to Fig. 1, it appears of interest to ask how the angle ζ changes depending on the domain model-specific properties of the EVD. Essentially, the angle ζ is a measure to which extent the E-value deviates from the "truth". Analytically, the relationship is given by Eq. (15) in Sec. 4; graphically, the relationship between the angle ζ and the EVD parameter λ is presented in Fig. 6 for the 185 cases found for the "hmmpfam" global-mode search. Inspection shows that a group of 161 domains (all having less than 110 positions) cluster at the upper right of the graph. By the way, all of our examples in Tables 1 and 2 are from those domains.

Strikingly, there is an outlier group of 24 domain models with $\lambda < 0.05$ and $\zeta < -15°$ (see Table 3). These are very long models (between 640 and 1,460 alignment positions) that clearly contain more than one actual domain. Score-wise, the most significant true hits of these models are strongly separated from their best matches in other sequences (see hmmpfam outputs in Supplementary File 6 and also the example in Supplementary File 1), mostly, without any sampling of the intermediate region (small λ means small spread in scores). It appears that, in these cases, the domain model essentially memorizes the seed alignment and cannot extrapolate into the sequence space. Due to the insufficient sampling, the EVD parameters cannot be well determined from the empirical distribution of best scores in the sequence database (i.e. the database does not provide a good estimate for random sequence

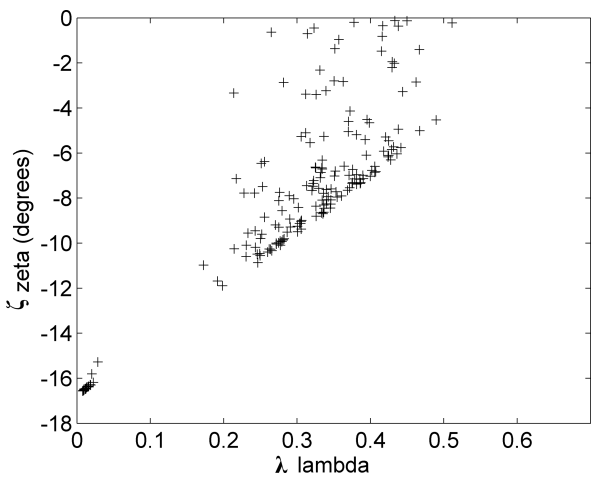

Fig. 6. Dependence of the angle ζ on the EVD parameter λ. The data involves 185 domain models that have a critical region in the hmmpfam global search mode. The analytical relationship is provided with Eq. (15) (Sec. 4). Twenty-four domain models with $\lambda < 0.05$ and $\zeta < -15°$ are clear outliers.

Table 3. Models with small $\lambda(\lambda < 0.05)$ and large negative $\zeta(\zeta < -15°)$.

Accession	Description
PF04730.4	Agrobacterium_VirD5_protein
PF02029.7	Caldesmon
PF00311.9	Phosphoenolpyruvate_carboxylase
PF02691.7	Vacuolating_cyotoxin
PF10433.1	Mono-functional_DNA-alkylating_agent_methyl_methanesulfonate
PF08377.2	MAP2_Tau_projection_domain
PF06933.3	Special_lobe-specific_silk_protein_SSP160
PF04094.6	Protein_of_unknown_function_DUF390
PF05567.3	Neisseria_PilC_protein
PF05483.4	Synaptonemal_complex_protein_1_SCP-1
PF03971.6	Monomeric_isocitrate_dehydrogenase
PF03344.7	Daxx_Family
PF07111.4	Alpha_helical_coiled-coil_rod_protein_HCR
PF05955.3	Equine_herpesvirus_glycoprotein_gp2
PF07218.3	Rhoptry-associated_protein_1_RAP-1
PF04147.4	Nop14-like_family
PF03157.5	High_molecular_weight_glutenin_subunit
PF03429.5	Major_surface_protein_1B
PF09731.1	Mitochondrial_inner_membrane_protein
PF02057.7	Glycosyl_hydrolase_family_59
PF05474.3	Semenogelin
PF04931.5	DNA_polymerase_phi
PF07217.3	Heterokaryon_incompatibility_protein_Het-C
PF06375.3	Bovine_leukaemia_virus_receptor_BLVR

matches) and, hence, a revision of the long domain models might be advisable in context of the global-mode search.

The results for the 1,748 domain models that were found in the "hmmsearch" global-mode search are similar. In total, 1,697 models have $\zeta > -15°$ and $\lambda > 0.05$. Their lengths are all less than 400 positions. The outlier group of 51 domains with $\zeta < -15°$ and $\lambda < 0.05$ has domain lengths between 640 and 1,547 positions.

In the Supplementary File 1, the global-mode "hmmsearch" result for Pfam domain model PF00311.9 (Phosphoenolpyruvate carboxylase with 960 positions) is shown. Its breakpoint occurs at score 11.7 with the corresponding E-value of 3.03. Beyond the breakpoint with increasing score, the E-value is returned by the logistic function as an exaggerated value. The top hit Q01647.1 with the score of 2796.7, the one with the largest score, has an EVD E-value of 8e-11 scaled all the way down to the E-value of 0 (a value smaller than the smallest represented positive number, certainly <1.e-300) by the logistic function. Thus, the change in order of magnitude is extremely big and solely caused by switching the statistical model.

On the other hand, we also have hits with EVD E-value >0.1 that show stretches of impressive sequence matches but deemed as insignificant under the EVD statistical model (for example, YP_003108745.1). Given the length of this particular

domain model, the global-mode can drastically under-sample random sequence matches and, thus, creates extremely small λ (<0.05). This leads to the underestimation of significance for those apparently good hits.

For these long domain models, the EVD statistical model can generally underestimate the significance of sequence matches. Yet ironically, it exaggerates the significance for hits with larger scores. In hindsight, the lack of sequence sampling in between the extremes of E-values is just an illusion that is created by the switching between two statistical models (EVD and logistic function).

3.5. *About the E-value computation in HMMER3*

The code of HMMER3 is completely new compared to HMMER2 and this includes also the E-value computing routines that are based solely on EVD-type functions. In our testing with a variety of domain models from Pfam release 24, we did not see any breakpoint as described in Fig. 1. Interestingly in HMMER3, the λ is predetermined by $\log 2 + \frac{1.44}{hN}$ where h is the average relative entropy per match state emission distribution, typically about 1.8 while N is the length of the query model, typically about 140.[22] Apparently, all Pfam models in release 24 have $\lambda > \log 2$. Indeed, the distribution of λ for Pfam release 24 is $\lambda \sim N(0.711, 0.0008)$ as calculated by us. It should be noted that this normal distribution is quite close to that of the fragment mode parameters in Pfam release 22 and 23 [$N(0.668, 0.051)$ for both cases] whereas the mean value for λ in the global mode is much lower. It remains open to which extent domain length and sequence space sampling by the database will influence the EVD parameters for the global mode (glocal as coined by Sean Eddy, p. 4 in Ref. 22) and whether a global mode option might be useful for further versions of HMMER.

The choice for an almost constant, high λ close to the magic value $\log 2$ for fragment mode models is based heavily on the work by Bundschuh, Milosavljević Yu, Hwa *et al.*[39–42] that both Viterbi scores of Gumbel distribution and Forward scores of exponential function has a fixed $\lambda = \log z$ where z is the base of the logarithm used of the log-odd scoring system. It should be noted that, in the case of $\lambda > \log 2$, the EVD is always below the respective logistic function and no switching would occur (Fig. 2).

Due to larger λ, the release 24 Pfam models are theoretically expected to have steeper E-value decrease for growing score (thus, higher sensitivity) than those from Pfam release 23 where most of the models (10,337 for global mode, 7,789 for fragment mode) have $\lambda < \log 2$. Ironically, the median gathering threshold for Pfam release 24 is lower than that of release 23 local models (21.4 versus 25), suggesting that the threshold needs to be lowered for comparable sensitivity. Also, it remains to be seen whether it is wise to adopt almost the same λ for the all Pfam release 24 models given that sequence space is usually biased. Sean Eddy himself has noted that some Pfam models (e.g. Ribosomal_L12 and XYPPX) have a λ that is significantly different from $\log 2$ (p. 7 in Ref. 22).

Time will provide arguments whether HMMER2 will finally be substituted by a version of HMMER3 and whether the list of discoveries made with HMMER3 can rival that of its predecessor. Since HMMER2 and HMMER3 generate different hit lists, it might become difficult to reproduce earlier publications, for example those on the evolutionary history of particular domains. The testing of the relative performance is a piece of laborious research that goes clearly beyond the scope of this article. Such a comparison has to verify whether differences in sensitivity between HMMER2 and HMMER3 are truly the result of changes in the respective theoretical frameworks or just a profane consequence of parameter adjustments such as gathering scores for the respective domain model releases. Regardless of the difficult theoretical questions associated with significance assessment of domain model hits, it seems that, so far, nothing can substitute for the experienced eye of a sequence-analytic researcher in evaluating the biological relevance of predictions before they are proposed for experimental follow-up.

4. Materials and Methods

4.1. *Derivation of the breakpoint point score for a domain model*

In the following, we rely on the derivation of significance criteria provided by Sean Eddy in his HMMER2 manual from October 2003 (pp. 47–50). According to the Bayes' theorem, the posterior probability of the null hypothesis N given the profile HMM (hidden Markov model) M and data D (the score of the sequence hit) is given as

$$P(N|D) = \frac{P(D|N)P(N)}{P(D|M)P(M) + P(D|N)P(N)}. \tag{1}$$

Assuming that prior probabilities are equal probable [i.e. $P(M) = P(N)$], the posterior null probability can be simplified to

$$P(N|D) = \frac{P(D|N)}{P(D|M)}P(M|D), \tag{2}$$

where $P(M|D) = 1 - P(N|D)$.

On the other hand, the probability of type I error for pair-wise alignment scores of a given HMM is denoted by $P(S \geq s)$, where s is the log odd score of a particular alignment. Hence, the posterior null probability is related to the type I error by

$$P(S \geq s) = \frac{P(D|N)}{P(D|M)}P(M|D). \tag{3}$$

The first part of the right side is a log-odd ratio that, following the HMM procedure, is calculated as

$$\frac{P(D|N)}{P(D|M)} = e^{-s \log 2}. \tag{4}$$

For $P(M|D)$, a sigmoid-type dependency between 0 and 1 is assumed and, given the form of Eq. (4), the simple arithmetic expression was selected by Sean Eddy as convenient:

$$P(M|D) = \frac{e^{s \log 2}}{1 + e^{s \log 2}}. \tag{5}$$

This results in a logistic-function-type expression for

$$P_{\text{logistic}}(S \geq s) = \frac{1}{1 + e^{s \log 2}}. \tag{6}$$

Alternatively, the significance can be estimated via an EVD. Indeed, if we assume the scores of an HMM positioned over all segments of a sequence being normally distributed, their maxima collected from all sequences in the database are extreme-value distributed.[43] In this case, the arithmetic form is the Gumbel (maximum) distribution and is given as

$$P_{\text{EVD}}(S \geq s) = 1 - e^{-e^{-\lambda(s-\mu)}} = \frac{e^{-\lambda(s-\mu)}}{1!} - \frac{e^{-2\lambda(s-\mu)}}{2!} + \frac{e^{-3\lambda(s-\mu)}}{3!} - \cdots \tag{7}$$

Both functions (6) and (7) are monotonous; yet, the logistic function approaches unity typically at smaller s than the EVD. Thus, there is a breakpoint $S_{\text{breakpoint}} = s'$ that occurs at $P_{\text{EVD}}(S \geq s') = P_{\text{logistic}}(S \geq s')$. Its numerical value s can be found by solving the equation

$$\frac{e^{-\lambda(s'-\mu)}}{1!} - \frac{e^{-2\lambda(s'-\mu)}}{2!} + \frac{e^{-3\lambda(s'-\mu)}}{3!} - \cdots = \frac{1}{1 + e^{s' \log 2}}.$$

For sufficiently large, positive values of s' (i.e. $s' \gg 0$ and $s' > \mu$), all terms except for the first one on the left side can be omitted (as well as the unity in the denominator on the right side) and the equation is approximated by

$$\frac{e^{-\lambda(s'-\mu)}}{1!} = \frac{1}{e^{s' \log 2}} \quad \text{and} \quad \frac{e^{\lambda\mu}}{e^{\lambda s'}} = \frac{1}{e^{s' \log 2}},$$

resulting in

$$s' = \frac{\lambda\mu}{\lambda - \log 2}. \tag{8}$$

For the remaining case of negative and sufficiently large values of s' (i.e. $s' \ll 0$), we derive the approximated equation for the breakpoint from the following form

$$1 - e^{-e^{-\lambda(s-\mu)}} = \frac{1}{1 + e^{s' \log 2}}. \tag{9}$$

Note that the expression $e^{s' \log 2}$ approaches zero with negative s' (its value is denoted by c below) and, hence, the right-hand side approaches one. Therefore, Eq. (9) can be approximated by

$$1 - e^{-e^{-\lambda(s-\mu)}} = (1 - c) \quad \text{and} \quad \lambda(s - \mu) = -\log(-\log(c)),$$

resulting in

$$s' = \mu - \frac{\log(-\log(c))}{\lambda}. \tag{10}$$

To prevent undefined expressions involving log in practical calculations with Eq. (10), c is set at 10^{-5}. Note that all the values μ in Pfam models, whether for global or fragment searches, are negative. Hence, Eq. (8) gives a positive result while Eq. (10) delivers a negative value for s' as expected.

In the routine "PValue" from the source file "Mathsupport.c" (see Supplementary File 2), HMMER2 selects the smaller (called P below) of the two values P_{logistic} and P_{EVD} for further computation of an E-value via

$$E(s) = nP(s), \tag{11}$$

where n is the database size. Thus, P-value calculation switches from the EVD to the logistic function for scores above $S_{\text{breakpoint}}$. Therefore, E-values decrease in a more pronounced manner with growing score than could be expected from the EVD (Fig. 1).

4.2. *Critical region of E-value ranges where insignificant EVD-derived E-values meet significant logistic function-derived E-values*

We calculated $S_{\text{breakpoint}}$ for the global search mode that forces complete domain model hits inside query sequences (the so-called ls-mode) both for the hmmpfam ($n =$ number of Pfam models $= 10{,}340$ for Pfam release 23[18,19]) and hmmsearch [$n =$ number of sequences in the non-redundant database $= 10{,}818{,}955$ (NR from 5th of April 2010)] HMMER2 applications.

Conclusions with regard to the validity of HMM hits in query sequences can be made in two ways. On the one hand, so-called gathering scores are applied that correspond to lowest score values of known true hits without false hits with higher score. In such a scheme, significance estimates are essentially not necessary. Alternatively, the goodness of HMM hit is evaluated with critical E-value thresholds of acceptable false-positive prediction. The E-value threshold recommended by Sean Eddy in his HMMER2 manual from October 2003 (p. 43) is 0.1. It should be noted that, by far, $S_{\text{breakpoint}}$ values for most HMM models in Pfam correspond to E-values clearly below 0.1 in any tested search HMMER2 regime. Thus, switching between the two statistical functions has no effect on the significance conclusion; yet, it adds to the cosmetics of the results by creating impressively low E-values for many good hits.

Unfortunately, there are some domain models for which $S_{\text{breakpoint}}$ corresponds to an E-value considerably larger than 0.1 (Fig. 1). In these cases, there is a critical score interval $(S_{\text{lower}}, S_{\text{upper}})$ excluding the boundary with the condition

$$\forall\, s \in (S_{\text{lower}}, S_{\text{upper}}) : E_{\text{EVD}}(s) > 0.1 \wedge E_{\text{logistic}}(s) < 0.1. \tag{12}$$

Thus, it can happen for some scores that an E-value calculated with EVD is not significant but that with the logistic function is.

The limiting score S_{upper} can be found by solving the EVD [Eq. (7)] for $\text{E} = 0.1$ as follows:

$$\frac{0.1}{n} = 1 - e^{-e^{-\lambda(S_{\text{upper}} - \mu)}}$$

and

$$S_{\text{upper}} = \mu - \frac{1}{\lambda} \log\left[-\log\left(1 - \frac{0.1}{n}\right)\right]. \tag{13}$$

Similarly, the lower limit score S_{lower} results as argument of the logistic function [Eq. (6)] for $\text{E} = 0.1$

$$\frac{0.1}{n} = \frac{1}{1 + e^{S_{\text{lower}} \log 2}}$$

and

$$S_{\text{lower}} = \frac{1}{\log 2}\left[\log\left(1 - \frac{0.1}{n}\right) - \log\left(\frac{0.1}{n}\right)\right] \tag{14}$$

With reference to Fig. 1, we can determine the angle ζ between the two straight lines representing the EVD and the logistic function at their crossing point $S_{\text{breakpoint}}$ via

$$\zeta = \arctan\left[\frac{\log(n(1 + \exp(S_{\text{upper}} \cdot \log 2))^{-1})}{S_{\text{upper}} - S_{\text{breakpoint}}}\right] - \arctan\left[\frac{-1}{S_{\text{upper}} - S_{\text{breakpoint}}}\right]. \tag{15}$$

Following Eqs. (8) and (13), ζ depends mainly on λ and is, as a trend, linearly related to it (Fig. 6).

Supplementary Materials

Additional data files for this article are also available via http://mendel.bii.a-star.edu.sg/SEQUENCES/ProblemDomains-JanusEvalue.

Supplementary File 1

Hmmpfam output of the Pfam domain PF00311.9 (Phosphoenolpyruvate carboxylase) when searched against the non-redundant database. We used the command "hmmsearch –Z 10,340" to annotate sequences with just this particular domain and to have E-values that equal to those searched with hmmpfam. The results contain true hits with extremely low E-values as well as good hits with large E-values. The E-value space in between the extremes is undersampled.

Supplementary File 2

Source code of "Mathsupport.c" in HMMER2 package with additional commentary lines.

Supplementary File 3

Detail information associated to the 185 and 1,748 domain models detected by global-mode hmmpfam and hmmsearch respectively.

Supplementary File 4

The concatenated HMMER2 outputs of the nine false-positive hits detected by global-mode hmmpfam as listed in Table 1 and illustrated in Fig. 4.

Supplementary File 5

The concatenated HMMER2 outputs of the seven false-positive hits detected by global-mode hmmsearch as listed in Table 2 and illustrated in Fig. 5.

Supplementary File 6

The archive (in WinRAR format) of the 24 outlier Pfam domains with small λ and large negative ζ listed in Table 3.

References

1. Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, Yuan Y, Predicting function: From genes to genomes and back, *J Mol Biol* **283**:707–725, 1998.
2. Devos D, Valencia A, Practical limits of function prediction, *Proteins* **41**:98–107, 2000.
3. Sander C, Schneider R, Database of homology-derived protein structures and the structural meaning of sequence alignment, *Proteins* **9**:56–68, 1991.
4. Todd AE, Orengo CA, Thornton JM, Evolution of function in protein superfamilies, from a structural perspective, *J Mol Biol* **307**:1113–1143, 2001.
5. Bork P, Gibson TJ, Applying motif and profile searches, *Methods Enzymol* **266**:162–184, 1996.
6. Durbin R, Eddy S, Krogh A, Mitchison G, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, 1999.
7. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, *Nucleic Acids Res* **25**:3389–3402, 1997.
8. Schneider G, Neuberger G, Wildpaner M, Tian S, Berezovsky I, Eisenhaber F, Application of a sensitive collection heuristic for very large protein families: Evolutionary relationship between adipose triglyceride lipase (ATGL) and classic mammalian lipases, *BMC Bioinformatics* **7**:164, 2006.
9. Henikoff JG, Greene EA, Taylor N, Henikoff S, Pietrokovski S, Using the blocks database to recognize functional domains, *Curr Protoc Bioinformatics*, Chapter 2: Unit 2.2, 2002.
10. Marchler-Bauer A, Anderson JB, Chitsaz F, Derbyshire MK, Weese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Liebert CA, Liu C, Lu F, Lu S, Marchler GH, Mullokandov M, Song JS, Tasneem A, Thanki N, Yamashita RA, Zhang D, Zhang N, Bryant SH, CDD: Specific functional annotation with the Conserved Domain Database, *Nucleic Acids Res* **37**:D205–D210, 2009.

11. Portugaly E, Linial N, Linial M, EVEREST: A collection of evolutionary conserved protein domains, *Nucleic Acids Res* **35**:D241–D246, 2007.
12. Schaffer AA, Wolf YI, Ponting CP, Koonin EV, Aravind L, Altschul SF, IMPALA: Matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices, *Bioinformatics* **15**:1000–1011, 1999.
13. Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, Rabkin S, Guo N, Muruganujan A, Doremieux O, Campbell MJ, Kitano H, Thomas PD, The PANTHER database of protein families, subfamilies, functions and pathways, *Nucleic Acids Res* **33**:D284–D288, 2005.
14. Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, Moulton G, Nordle A, Paine K, Taylor P, Uddin A, Zygouri C, PRINTS and its automatic supplement, prePRINTS, *Nucleic Acids Res* **31**:400–402, 2003.
15. Bru C, Courcelle E, Carrere S, Beausse Y, Dalmar S, Kahn D, The ProDom database of protein domain families: More emphasis on 3D, *Nucleic Acids Res* **33**:D212–D215, 2005.
16. Hulo N, Bairoch A, Bulliard V, Cerutti L, Cuche BA, de CE, Lachaize C, Langendijk-Genevaux PS, Sigrist CJ, The 20 years of PROSITE, *Nucleic Acids Res* **36**:D245–D249, 2008.
17. Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, Madera M, Chothia C, Gough J, SUPERFAMILY — sophisticated comparative genomics, data mining, visualization and phylogeny, *Nucleic Acids Res* **37**:D380–D386, 2009.
18. Bateman A, Birney E, Durbin R, Eddy SR, Finn RD, Sonnhammer EL, Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins, *Nucleic Acids Res* **27**:260–262, 1999.
19. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL, The Pfam protein families database, *Nucleic Acids Res* **28**:263–266, 2000.
20. Letunic I, Doerks T, Bork P, SMART 6: Recent updates and new developments, *Nucleic Acids Res* **37**:D229–D232, 2009.
21. Eddy SR, What is a hidden Markov model? *Nat Biotechnol* **22**:1315–1316, 2004.
22. Eddy SR, A probabilistic model of local sequence alignment that simplifies statistical significance estimation, *PLoS Comput Biol* **4**:e1000069, 2008.
23. Eddy SR, A new generation of homology search tools based on probabilistic inference, *Genome Inform* **23**:205–211, 2009.
24. Gonzalez MW, Pearson WR, RefProtDom: A protein database with improved domain boundaries and homology relationships, *Bioinformatics* **26**:2361–2362, 2010.
25. Kann MG, Sheetlin SL, Park Y, Bryant SH, Spouge JL, The identification of complete domains within protein sequences using accurate E-values for semi-global alignment, *Nucleic Acids Res* **35**:4678–4685, 2007.
26. Wong WC, Maurer-Stroh S, Eisenhaber F, More than 1,001 problems with protein domain databases: Transmembrane regions, signal peptides and the issue of sequence homology, *PLoS Comput Biol* **6**:e1000867, 2010.
27. Mott R, Accurate formula for P-values of gapped local sequence and profile alignments, *J Mol Biol* **300**:649–659, 2000.
28. Waterman MS, Vingron M, Rapid and accurate estimates of statistical significance for sequence data base searches, *Proc Natl Acad Sci USA* **91**:4625–4628, 1994.
29. Agrawal A, Brendel VP, Huang X, Pairwise statistical significance and empirical determination of effective gap opening penalties for protein local sequence alignment, *Int J Comput Biol Drug Des* **1**:347–367, 2008.
30. Karlin S, Dembo A, Kawabata T, Statistical composition of high-scoring segments from molecular sequences, *The Annals of Statistics* **18**:571–581, 1990.

31. Karlin S, Altschul SF, Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes, *Proc Natl Acad Sci USA* **87**:2264–2268, 1990.
32. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, Basic local alignment search tool, *J Mol Biol* **215**:403–410, 1990.
33. Altschul SF, Wootton JC, Gertz EM, Agarwala R, Morgulis A, Schaffer AA, Yu YK, Protein database searches using compositionally adjusted substitution matrices, *FEBS J* **272**:5101–5109, 2005.
34. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF, Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements, *Nucleic Acids Res* **29**:2994–3005, 2001.
35. Yu YK, Wootton JC, Altschul SF, The compositional adjustment of amino acid substitution matrices, *Proc Natl Acad Sci USA* **100**:15688–15693, 2003.
36. Yu YK, Altschul SF, The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions, *Bioinformatics* **21**:902–911, 2005.
37. Yu YK, Gertz EM, Agarwala R, Schaffer AA, Altschul SF, Retrieval accuracy, statistical significance and compositional similarity in protein sequence database searches, *Nucleic Acids Res* **34**:5966–5973, 2006.
38. Nadarajah S, Linear combination of Gumbel random variables, *Stoch Environ Res Risk Assess* **21**:283–286, 2007.
39. Milosavljević A, Jurka J, Discovering simple DNA sequences by the algorithmic significance method, *Comput Appl Biosci* **9**:407–411, 1993.
40. Yu YK, Bundschuh R, Hwa T, Hybrid alignment: High-performance with universal statistics, *Bioinformatics* **18**:864–872, 2002.
41. Bundschuh R, Rapid significance estimation in local sequence alignment with gaps, *J Comput Biol* **9**:243–260, 2002.
42. Chia N, Bundschuh R, A practical approach to significance assessment in alignment with gaps, *J Comput Biol* **13**:429–441, 2006.
43. Kotz S, Nadarajah S, *Extreme Value Distributions Theory and Applications*, Imperial College Press, London, 2000.
44. Kamiya N, Shen JR, Crystal structure of oxygen-evolving photosystem II from Thermosynechococcus vulcanus at 3.7-A resolution, *Proc Natl Acad Sci USA* **100**:98–103, 2003.
45. Walters KJ, Lech PJ, Goh AM, Wang Q, Howley PM, DNA-repair protein hHR23a alters its protein structure upon binding proteasomal subunit S5a, *Proc Natl Acad Sci USA* **100**:12694–12699, 2003.
46. Gomis-Ruth FX, Sola M, Acebo P, Parraga A, Guasch A, Eritja R, Gonzalez A, Espinosa M, del SG, Coll M, The structure of plasmid-encoded transcriptional repressor CopG unliganded and bound to its operator, *EMBO J* **17**:7404–7415, 1998.
47. Isaza CE, Silaghi-Dumitrescu R, Iyer RB, Kurtz DM, Jr., Chan MK, Structural basis for O2 sensing by the hemerythrin-like domain of a bacterial chemotaxis protein: Substrate tunnel and fluxional N terminus, *Biochemistry* **45**:9023–9031, 2006.
48. Yoshida A, Tomita T, Kurihara T, Fushinobu S, Kuzuyama T, Nishiyama M, Structural insight into concerted inhibition of alpha 2 beta 2-type aspartate kinase from Corynebacterium glutamicum, *J Mol Biol* **368**:521–536, 2007.
49. Cheng YS, Hsia KC, Doudeva LG, Chak KF, Yuan HS, The crystal structure of the nuclease domain of colicin E7 suggests a mechanism for binding to double-stranded DNA by the H-N-H endonucleases, *J Mol Biol* **324**:227–236, 2002.

50. Nonaka T, Fujihashi M, Kita A, Saeki K, Ito S, Horikoshi K, Miki K, The crystal structure of an oxidatively stable subtilisin-like alkaline serine protease, KP-43, with a C-terminal beta-barrel domain, *J Biol Chem* **279**:47344–47351, 2004.
51. Zhang Z, Zhu L, Lin D, Chen F, Chen DJ, Chen Y, The three-dimensional structure of the C-terminal DNA-binding domain of human Ku70, *J Biol Chem* **276**:38231–38236, 2001.
52. Julien O, Gignac I, Hutton A, Yee A, Arrowsmith CH, Gagne SM, MTH187 from Methanobacterium thermoautotrophicum has three HEAT-like repeats, *J Biomol NMR* **35**:149–154, 2006.
53. Korotkov KV, Pardon E, Steyaert J, Hol WG, Crystal structure of the N-terminal domain of the secretin GspD from ETEC determined with the assistance of a nanobody, *Structure* **17**:255–265, 2009.
54. Bogden CE, Fass D, Bergman N, Nichols MD, Berger JM, The structural basis for terminator recognition by the Rho transcription termination factor, *Mol Cell* **3**:487–493, 1999.
55. Tanaka Y, Sakamoto S, Kuroda M, Goda S, Gao YG, Tsumoto K, Hiragi Y, Yao M, Watanabe N, Ohta T, Tanaka I, A helical string of alternately connected three-helix bundles for the cell wall-associated adhesion protein Ebh from Staphylococcus aureus, *Structure* **16**:488–496, 2008.

Wing-Cheong Wong graduated with a B.Sc. (Hons) in Computer Engineering from Nanyang Technological University (NTU), Singapore, in 2002. In 2004, he obtained his M.Sc. in Bioinformatics from the Faculty of Medicine at National University of Singapore (NUS) under a scholarship program awarded by the Bioinformatics Institute (BII) of Agency for Science, Technology and Research (A*STAR). He is currently a senior research associate with the Protein Sequence Analysis Group in the Bioinformatics Institute. His area of research in computational biology encompasses nucleic/protein sequence studies as well as expression profile analysis. His strength is in the application of rigorous statistical methods on biological problems.

Sebastian Maurer-Stroh studied theoretical biochemistry in the group of Peter Schuster at the University of Vienna and wrote his master's and Ph.D. thesis on problems of posttranslational modification prediction from protein sequences in Frank Eisenhaber's group at the Institute of Molecular Pathology (IMP) in Vienna. After FEBS and Marie Curie fellowships at the VIB-SWITCH lab in Brussels, he joined the A*STAR Bioinformatics Institute (BII), Singapore, where he is leading the Protein Sequence Analysis Group since 2007. He has contributed widely used predictors for posttranslational lipid modifications, amyloid fiber formation and catalyzed new biomolecular insights by sequence-based function predictions. Being at the forefront of research during the swine flu pandemic, he is also coordinating the institute's cross-divisional program "Human Infectious Diseases."

Frank Eisenhaber studied mathematics at the Humboldt-University in Berlin and biophysics and medicine at the Pirogov Medical University in Moscow. He was awarded a M.D. in 1985. Three years later, he received the Ph.D. in Molecular Biology from the Engelhardt Institute of Molecular Biology in Moscow (supervision by Dr. Vladimir Gayevich Tumanyan). After post-doctoral work at the Institute of Molecular Biology in Berlin-Buch (1989–1991) and at the EMBL in Heidelberg (1991–1999), he worked as team leader of the bioinformatics research group and head of the general IT department at the Institute of Molecular Pathology (IMP) in Vienna (1999–2007). Since August 2007, he is the Director of the Bioinformatics Institute A*STAR Singapore. Frank Eisenhaber's research interest is focused on the discovery of new biomolecular mechanisms with theoretical and biochemical approaches and the functional characterization of yet uncharacterized genes and pathways. Frank Eisenhaber is one of the scientists credited with the discovery of the SET domain methyltransferases, ATGL, kleisins, many new protein domain functions and with the development of accurate prediction tools for posttranslational modifications and subcellular localizations.