Biđin *et al*., 2020; **SUPPLEMENTARY**

1. *Leitmotif's* architecture and parameterizations

1.1. *Leitmotif's* architecture

Motif-HMM was first described by Grundy *et al*. (1997). The original motif-HMM (mHMM) has a sequence of *match states* (without *insert/delete states*), flanked by single self-looping *insert states* that model inter-motif regions. *Leitmotif* differs from the original version somewhat since it models the motif region with a profile-HMM type architecture (described for example in Durbin *et al.*, 1998). The advantage of this architecture is that it allows insertions and deletions in the motif. Accordingly, *Leitmotif* has the **TP** (Transition Probability) parameter which corresponds to the *match* to *match* transition probability in the motif region (e.g. $P^i_{M_1M_2}$ in Fig. 1). If the value of this parameter is set to less than one, this will allow *match* to *insert* (i.e. insertion) and *match* to *delete* (i.e. deletion) transitions.

A motif region of the mHMM (e.g. *Leitmotif*) is much shorter than the typical profile-HMM and its length equals the motif length. Emission probabilities of the self-looping *insert states* in *Leitmotif* (shaded grey in Fig.1) are set to background probabilities. Since the sequence score is computed as a log-likelihood ratio of the given model over the random model (with the random model having background emission probabilities) amino acid residues of the inter-motif regions have no contribution to the final score. This is desirable since in contrast to profile-HMM, in motif scanning one is interested only in finding motifs; inter-motif regions (residues) are of no interest. Transition Probabilities (**TP**) of the inter-motif self-looping *insert states* are computed based on the expected inter-motif distance provided by the user. Namely, in accordance with the model architecture, the random variable corresponding to the number of inter-motif residues follows the geometric distribution (with support $R_X = \mathbb{N}$ and $E(X) = 1/P_i^R$). For example, with an inter-motif expected distance of 100 residues, it easily follows that $P_i^R = 0.01$.

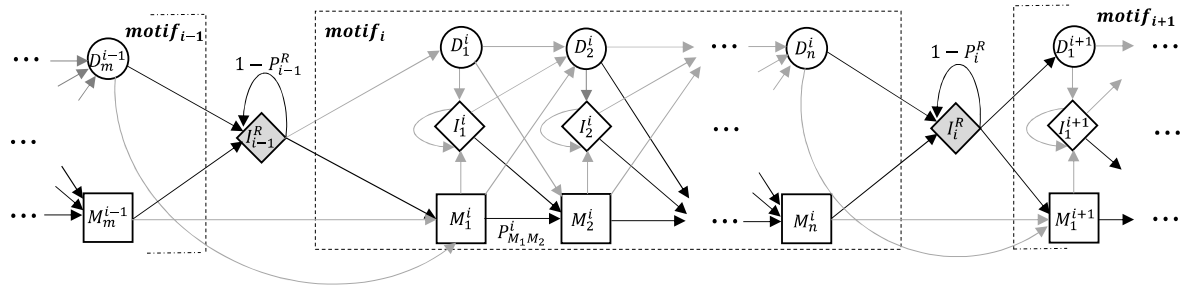


Fig. 1. ***Leitmotif's*** **architecture**. A motif region (dashed box) has the same architecture as the standard profile-HMM. Unlike the profile-HMM, a motif region is in general very short. Each "column" in a motif region is composed of stacked *match*, *insert* and *delete states*. For example, if a motif is 10 residues long, the corresponding motif region will be composed of 10 *match-insert-delete* columns. The figure above shows two *match-insert-delete* columns followed by an unspecified number of columns represented by three dots and ending in the final column. Note that the final column lacks an *insert* state. Motif regions are followed by single self-looping *insert states* (shaded in grey) which model regions between motifs, and regions between a motif and the N/C terminus. Their emission probabilities are set to background probabilities. Transition probabilities of the inter-motif *insert states* are computed based on user-defined expected inter-motif distances as described above.

1.2 A heuristic formula for determining distance penalties

$$final\_seq\_score = seq\_score - k \cdot maxscore \cdot \log_{10}(|D| + 1)$$

seq_score is the Viterbi log-likelihood ratio for a given sequence; $|D|$ is the absolute value of the difference between the user-specified expected distance between motifs (or between a motif and the N/C terminus) and the distance returned by the algorithm. $maxscore$ is the Viterbi log-likelihood score of the optimal sequence (i.e. artificial sequence containing the most frequent motif residues and having user-defined motif distances). This term is necessary since sequence scores vary significantly depending on the number of given motifs as well as motif length(s). k is a parameter which depends on the user-defined penalty strength. Namely, $k = 0$ for no penalty, $k = 0.1$ for Weak penalty, $k = 1$ for Medium penalty and $k = 10$ for Strong penalty. Therefore penalty increases incrementally by one order of magnitude with each step. Individual distance penalties are averaged and subtracted from the sequence score.

1.3. Computing emission probabilities and sequence weighting options

As stated in the manuscript, *Leitmotif* offers four different ways of computing *match state* emission probabilities. Different algorithms were implemented since: (**i**) HMMER uses Dirichlet Mixture (**DM**) as default for its *hmmsearch* program with accompanying priors as described in (Sjölander *et al*., 1996); (**ii**) Henikoff Algorithm (**HA**) showed superior performance in (Henikoff *et al*., 1996). Note that only the HA uses substitution matrices; (**iii**) Modified Ancestral algorithm (**MA**) was used in our previous study (Vujaklija *et al*., 2016) where it showed very good performance; and (**iv**) Simple Pseudocounts algorithm (**SP**; Durbin *et al*., 1998) is appropriate in cases of large and phylogenetically representative seed.
Leitmotif uses sequence weighting by default since it is a standardly used method to reduce the influence of phylogenetically biased samples (Durbin *et al*., 1998). Note that only the **DM** algorithm features both relative and total sequence weighting since it is the only algorithm using priors.

2. Algorithm performance on different datasets

In the Results section of our manuscript we summarized *Leitmotif's* performance on two datasets (the GDSL protein family and the Desaturases First subfamily) using selected **IR/DP** parameters (Table 1 A&B in the main text). Detailed results of these analyses are described below.

2.1 ROC scores

Leitmotif performances were compared using ROC curves, ROC scores and slightly modified versions of the ROC n scores which we call a normalised ROC n or nROC n for short (Grundy *et al*., 1997 used the same normalisation). The ROC n is defined as the area under the ROC curve until the first n negatives are found (Gribskov, M. and Robinson, N.L. 1996). The advantage of ROC n scores over standard ROC is described in detail in (Gribskov, M. and Robinson, N.L. 1996).
The problem with ROC n however, is that it sometimes doesn't give very intuitive results. For example, if there are 100 negatives in the test dataset and an algorithm yields a perfect ranking with all positives ranked above all the negatives (thus the ROC curve is a stepwise function $f(x) = 1;\ x \in [0,1]$), the ROC 1 score is just 1/100 = 0.01 (i.e. the area under the ROC curve corresponding to a TN rate of 0.01). In case of a perfect ranking with 500 negatives ROC 1 would be 1/500 = 0.002 (i.e. the area under the ROC curve corresponding to a TN rate of 0.002), although the ranking is ideal in both cases. Hence, it makes sense to normalize this number (ROC n) by dividing it by the area under an ideal stepwise ROC curve up to the corresponding percentage of true negatives (i.e. n divided by the total number of negatives). In the example above that would be nROC 1 = (ROC 1)/0.01 = 1 for the 100 negatives case

and analogously nROC 1 = (ROC 1)/0.002 = 1 for the 500 negatives case, thus giving the more intuitive result. Note that both ROC n and nROC n (unlike standard ROC) are meaningful as a comparison measure of different algorithms only on the *same test dataset* since both depend on the number of negatives in the test dataset.

2.2 GDSL family analysis

In the Results section in the main text (Table 1A) we summarized results of the GDSL family analysis using selected **IR/DP** parameter values. The GDSL dataset used in this study is almost identical to the test dataset used in Vujaklija I *et al*., (2016). It comprises 804 GDSL sequences (Supplement GDSL Test sequences & GDSL Sequence IDs) divided into two subsets, 624 Positives and 180 Negatives. It was obtained by removing, from the initial GDSL dataset (Vujaklija, I *et al*., 2016), 16/820 sequences (2%) since they were either identical, highly similar ($\geq$ 98%) or very ambiguous and we were uncertain how to assign them, to Positives or Negatives. The removal of these sequences did not change significantly the ratio of Positives vs Negatives. It was changed from 20.85% to 21.89%.

2.2.1. Selection of DP/IR parameters

The reasons for using Weak (W), Medium (M) and Strong (S) Distance Penalties (**DP**) in cases of one, two and three GDSL motifs respectively are described below.

The **DP** parameter's usefulness will strongly depend on the protein family being analysed. In some cases where distances are (very) conserved this parameter is very useful (e.g. Desaturases First subfamily described in 2.3 is an excellent example). In cases where distances are moderately conserved, the usefulness of this parameter will be moderate. In other cases where motif distances are not conserved (i.e. highly variable) the effect of DP will be insignificant or even detrimental.

In the case of GDSL lipases, it is known from literature that their distances are moderately conserved. Thus, the presence of all three conserved motifs is a much stronger predictor of family membership than motif distances. Moreover, setting **DP** to Strong (or Medium) would mean that one expects positives to have all four distances (two between motifs and two between a motif and N/C terminus) close to the expected distances, which is unlikely in GDSL case since motif distances are not strongly conserved. Accordingly we used a Weak **DP** in this case.
In case of two motifs the amount of useful information is decreased thus reducing *Leitmotif's* discriminatory power since it scans for the presence of only 2/3 motifs. Taking this into consideration our rationale was to try to compensate this negative effect by giving a stronger influence to **DP**. Namely we took into consideration that the GDSL catalytic domain is moderately conserved (around 400 residues) and that the 1st and 3rd motifs are located close to the N and C terminus respectively. Accordingly we used a Medium **DP** in this case.
Finally, in the case of only one motif, scanning is even more challenging. Analogously, we used the same rationale as in the two motifs case and thus have further increased **DP** to Strong.

Residues selected as immutable (**IR**) were Ser (1st motif) & His (3rd motif). This was based on reported literature data (Leščić Ašler *et al*., 2017) and our experimental results. Namely, the site-directed mutagenesis of several conserved residues confirmed the complete loss of enzyme activity only in the absence of Ser-His (A. Bielen PhD thesis, 2011, University of Zagreb). Our result was in line with the authors proposing that the catalytic mechanism is based on the Ser-His dyad.
Note that in the Desaturases First subfamily case, **IR** were selected exclusively based on the *seed dataset* (2.3.3)

2.2.2. Results of GDSL family analysis

Tables A1, A2 and A3: Different emission probability algorithms with/without IR/DP parameters

A1	ALG	IR	SW	TP	SM	MD	DP	ROC	nROC 50	nROC 5	nROC 1
3 MOTIFS	MA	[6S,,4H]	R	0.99	None	50:125:175:50	W	0.9902	0.9651	0.7968	0.7324
		[6S,,4H]					/	0.9895	0.9629	0.7936	0.6987
		[,,]					W	0.9800	0.9286	0.7314	0.609
		[,,]					/	0.9784	0.9239	0.7202	0.5994
	DM	[6S,,4H]	R&T	0.99	None	50:125:175:50	W	0.9899	0.9663	0.891	0.8798
		[6S,,4H]					/	0.9896	0.9651	0.8907	0.8734
		[,,]					W	0.9842	0.9481	0.8689	0.8333
		[,,]					/	0.9834	0.9457	0.8663	0.8349
		[6S,,4H]	R	0.99	None	50:125:175:50	W	0.9868	0.9536	0.8029	0.7115
		[6S,,4H]					/	0.9856	0.9495	0.7949	0.6891
		[,,]					W	0.9793	0.9269	0.8029	0.7115
		[,,]					/	0.9777	0.922	0.7917	0.6891
	HA	[6S,,4H]	R	0.99	BLOSUM 62	50:125:175:50	W	0.9883	0.9578	0.7814	0.75
		[6S,,4H]					/	0.9879	0.9563	0.7827	0.7388
		[,,]					W	0.9819	0.9361	0.7462	0.6603
		[,,]					/	0.9813	0.9338	0.7426	0.6426
	SP	[6S,,4H]	R	0.99	None	50:125:175:50	W	0.9832	0.9403	0.7372	0.7035
		[6S,,4H]					/	0.9832	0.9405	0.7423	0.6939
		[,,]					W	0.9782	0.9243	0.7372	0.7051
		[,,]					/	0.9778	0.9242	0.7285	0.6939

A2	ALG	IR	SW	TP	SM	MD	DP	ROC	nROC 50	nROC 5	nROC 1
2 MOTIFS	MA	[6S,4H]	R	0.99	None	50:300:50	M	0.9879	0.9588	0.8458	0.7901
		[6S,4H]					/	0.9774	0.9225	0.5125	0.2564
		[,]					M	0.9735	0.9233	0.7686	0.7196
		[,]					/	0.9665	0.9009	0.4965	0.2564
	DM	[6S,4H]	R&T	0.99	None	50:300:50	M	0.9807	0.9403	0.7186	0.5192
		[6S,4H]					/	0.9715	0.9056	0.4503	0.1282
		[,]					M	0.9734	0.9271	0.7109	0.5513
		[,]					/	0.9651	0.8936	0.4503	0.1282
		[6S,4H]	R	0.99	None	50:300:50	M	0.9865	0.9553	0.8295	0.7772
		[6S,4H]					/	0.9732	0.912	0.499	0.3109
		[,]					M	0.9774	0.9329	0.7939	0.7772
		[,]					/	0.9662	0.9038	0.499	0.3109
	HA	[6S,4H]	R	0.99	BLOSUM 62	50:300:50	M	0.985	0.954	0.833	0.7997
		[6S,4H]					/	0.9743	0.9125	0.4769	0.1138
		[,]					M	0.9753	0.9305	0.8006	0.7804
		[,]					/	0.9677	0.9021	0.475	0.1138
	SP	[6S,4H]	R	0,99	None	50:300:50	M	0,9813	0,9474	0,8394	0,8093
		[6S,4H]					/	0,9673	0,8988	0,4795	0,2228
		[,]					M	0,9711	0,9202	0,7635	0,6843
		[,]					/	0,9622	0,8924	0,4795	0,2228

A3	ALG	IR	SW	TP	SM	MD	DP	ROC	nROC 50	nROC 5	nROC 1
1 MOTIF	MA	[6S]	R	0.99	None	50:368	S	0.8916	0.6661	0.1083	0.008
		[6S]					/	0.7633	0.253	0.0144	0.0048
		[]					S	0.8601	0.5706	0.0154	0.008
		[]					/	0.7517	0.2483	0.0144	0.0048
	DM	[6S]	R&T	0.99	None	50:368	S	0.8991	0.6729	0.1622	0.0064
		[6S]					/	0.7596	0.2432	0.0042	0
		[]					S	0.7681	0.2754	0.0122	0.0016
		[]					/	0.7565	0.2432	0.0042	0
		[6S]	R	0.99	None	50:368	S	0.8945	0.6807	0.1583	0.008
		[6S]					/	0.7679	0.2701	0.0317	0.008
		[]					S	0.859	0.5699	0.0192	0.008
		[]					/	0.7625	0.2701	0.0317	0.008
	HA	[6S]	R	0.99	BLOSUM 62	50:368	S	0.8915	0.6693	0.109	0.008
		[6S]					/	0.7619	0.2563	0.0157	0.0112
		[]					S	0.8666	0.5856	0.0196	0.008
		[]					/	0.7579	0.2563	0.0157	0.0112
	SP	[6S]	R	0.99	None	50:368	S	0.8958	0.6857	0.109	0.008
		[6S]					/	0.765	0.2657	0.0224	0.0064
		[]					S	0.8704	0.6048	0.0215	0.008
		[]					/	0.7616	0.2657	0.0224	0.0064

ALG-Algorithms: **MA**-Modified Ancestral, **DM**-Dirichlet Mixture, **HA**-Henikoff Algorithm, **SP**-Simple Pseudocounts. **IR**-Immutable Residues. Motif descriptions: "[6S, ,4H]", "[6S,4H]", "[6S]". "6S" represents Ser at position 6 (motif 1); 4H represents His at position 4 (motif 3) (Fig. 1, upper panel in the main text). Motifs are separated by commas. **SW**-Sequence Weighting (**R**-Relative; **R**&**T**- Relative and Total), **TP**-*match* to *match* Transition Probability, **SM**-Substitution Matrix, **MD**- Motif Distances, **MD** were chosen based on reported data and expert knowledge (Vujaklija et al., 2016; Upton and Buckley, 1995). **DP**-Distance Penalty ("/"-None; "W"-Weak, "M"-Medium, "S"-Strong) were chosen as described in 2.2.1; Light grey rows show ROC scores with selected **IR** and **DP**; Dark grey rows show ROC scores without **IR/DP** parameters.

Tables A4, A5 and A6: Results of *Leitmotif* performance with Weak, Medium and Strong Distance Penalty (DP) for three, two and one motif.

A4	ALG	IR	SW	TP	SM	MD	DP	ROC	nROC 50	nROC 5	nROC 1
3 MOTIFS	MA	[6S,,4H]	R	0,99	None	50:125:175:50	W	0.9902	0.9651	0.7968	0.7324
		[6S,,4H]					/	0.9895	0.9629	0.7936	0.6987
		[6S,,4H]					M	0.9818	0.9348	0.6603	0.4663
		[6S,,4H]					S	0.9356	0.8012	0.2494	0.1939
		[,,]					W	0.9800	0.9286	0.7314	0.6090
		[,,]					/	0.9784	0.9239	0.7202	0.5994
		[,,]					M	0.9692	0.8955	0.6321	0.4696
		[,,]					S	0.9055	0.7062	0.2144	0.1779

A5	ALG	IR	SW	TP	SM	MD	DP	ROC	nROC 50	nROC 5	nROC 1
2 MOTIFS	MA	[6S,4H]	R	0,99	None	50:300:50	M	0.9879	0.9588	0.8458	0.7901
		[6S,4H]					W	0.9797	0.9303	0.5625	0.3333
		[6S,4H]					/	0.9774	0.9225	0.5125	0.2564
		[6S,4H]					S	0.9471	0.8284	0.4792	0.3429

		[,]					M	0.9735	0.9233	0.7686	0.7196
		[,]					W	0.9693	0.9096	0.5426	0.3285
		[,]					/	0.9665	0.9009	0.4965	0.2564
		[,]					S	0.9136	0.7245	0.1660	0.0256

A6	ALG	IR	SW	TP	SM	MD	DP	ROC	nROC 50	nROC 5	nROC 1
1 MOTIFS	MA	[6S]	R	0,99	None	50:368	S	0.8916	0.6661	0.1083	0.0080
		[6S]					M	0.8432	0.4881	0.1205	0.0433
		[6S]					W	0.7727	0.2771	0.0337	0.0112
		[6S]					/	0.7633	0.2530	0.0144	0.0048
		[]					S	0.8601	0.5706	0.0154	0.0080
		[]					M	0.8198	0.4399	0.1013	0.0497
		[]					W	0.7600	0.2713	0.0311	0.0112
		[]					/	0.7517	0.2483	0.0144	0.0048

ALG-Algorithms: **MA**-Modified Ancestral, Motif descriptions: "[6S, ,4H]", "[6S,4H]", "[6S]". "6S" represents Ser at position 6 (motif 1); 4H represents His at position 4 (motif 3) (Fig 1, upper panel in the main text). Motifs are separated by commas. **SW**-Sequence Weighting (**R**-Relative), **TP**-*match* to *match* Transition Probability, **SM**-Substitution Matrices, **MD**-Motif Distances. **MD** were chosen based on expert knowledge and reported data (Vujaklija et al., 2016; Upton and Buckley, 1995). **DP**-Distance Penalty ("/"-None; "W"-Weak, "M"-Medium, "S"-Strong); Light grey row shows ROC scores with selected **IR** and **DP**; Dark grey rows show ROC scores without **IR/DP** parameters.

2.2.3. ROC curves with different parameterizations

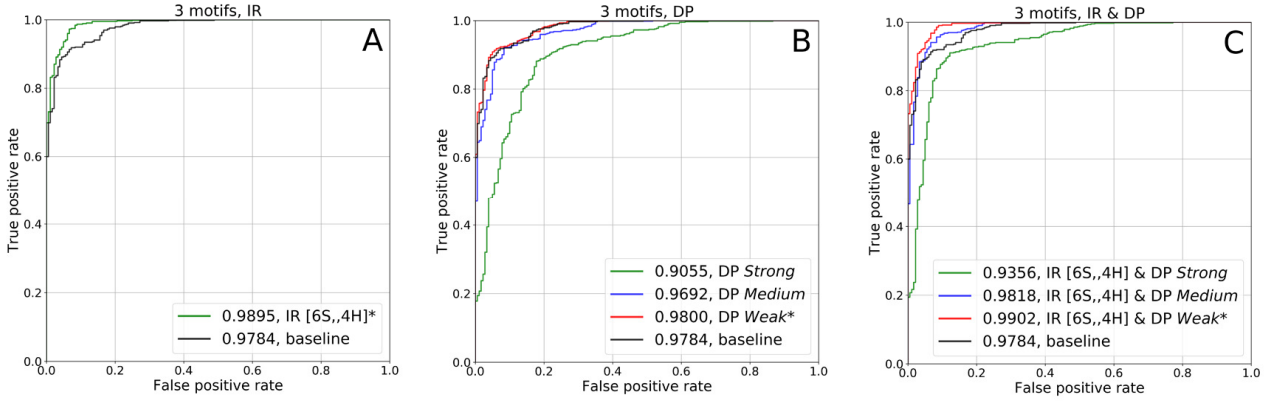


Fig. 2. ROC curves for three motifs - GDSL lipases dataset. Selected values for IR/DP are marked by asterisks. **A**) Results with Immutable Residues (**IR**) only. As shown in the figure setting **IR** improves the ROC score (i.e. area under the ROC curve)**. B**) Results with Distance Penalties (**DP**) only. As can be seen in the figure modest improvement in ranking is obtained by using Weak **DP. C**) Results of combining **DP** and **IR**. As shown in figures A, B, and C the best overall ROC score in this case is obtained by using a combination of Weak **DP** and **IR** (for explanation see 2.2.1).

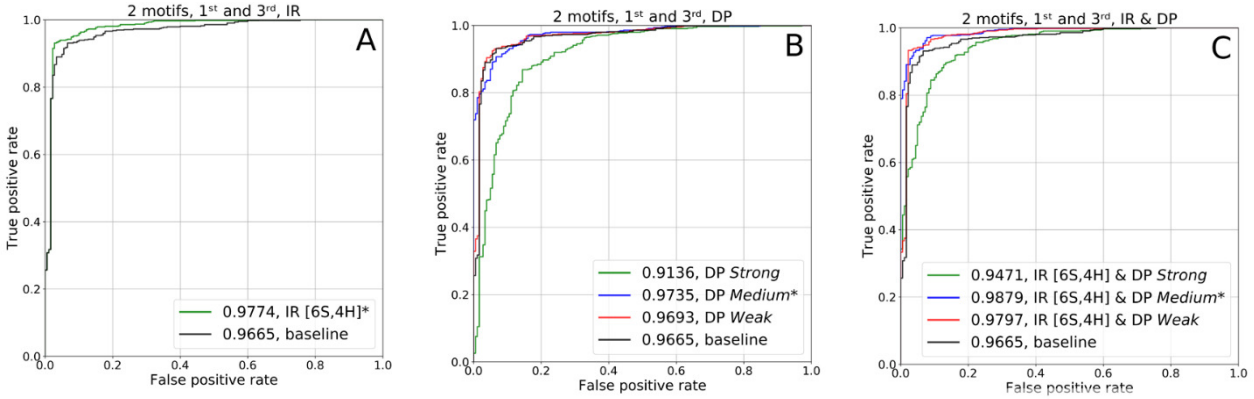


Fig. 3. ROC curves for two motifs - GDSL lipases dataset. Selected values for IR/DP are marked by asterisks. **A**) Results with Immutable Residues (**IR**) only. As shown in the figure setting **IR** improves the ROC score**. B**) Results with Distance Penalties (**DP**) only. As can be seen in the figure improvement in ranking is obtained by using Medium **DP. C**) Results of combining **DP** and **IR**. As shown in figures A, B, and C the best overall ROC score in this case is obtained by using a combination of Medium **DP** and **IR** (for explanation see 2.2.1).

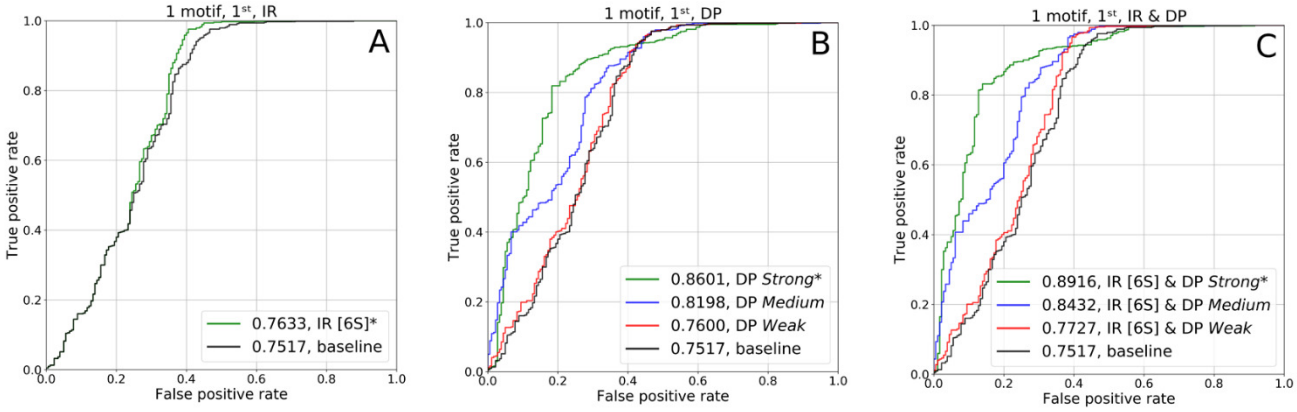


Fig. 4. ROC curves for one motif - GDSL lipases dataset. Selected values for IR/DP are marked by asterisks. **A**) Results with Immutable Residues (**IR**) only. As shown in the figure setting **IR** improves the ROC score**. B**) Results with Distance Penalties (**DP**) only. As can be seen in the figure significant improvement in ROC score is obtained by using Strong **DP. C**) Results of combining **DP** and **IR**. As shown in figures A, B, and C the best overall ROC score is obtained by using a combination of Strong **DP** and **IR** (for explanation see 2.2.1).

Altogether, results in (2.2.2. Tables A1-A6 & 2.2.3. Fig. 2-4) show best results (best ROC scores) are obtained when using all three motifs. This is expected since three motifs provide more information than one or two. In addition, using selected **IR** and **DP** (Weak, Medium and Strong for three, two and one motif respectively) marked by asterisks in Fig. 2-4 and shaded light grey in Tables A1-A6 increases ROC scores. The reasons for selecting these parameters are described in 2.2.1. For comparison purposes we also show ROC scores and ROC curves for all other **DP** values (Tables A1-A6; Fig. 2-4/B&C).

Next, the effect of distance penalties (/, W, M, S) was analysed in more detail. We examined the distances between motifs in positives and negatives in the whole GDSL dataset. Interestingly, there is little difference between positives and negatives in terms of mean and standard deviation for distances from the N-terminus to the 1st motif and moderate difference for distances from the 3rd motif to the C-terminus (Table A7). However, there is a large difference both in standard deviation (SD) and mean for the other two distances (from 1st to 2nd and from 2nd to 3rd motif).

Table A7 Variations of motif distances in the GDSL dataset (Mean ± SD)

Distances from N-terminus to 1st motif		**Distances from 1st motif to 2nd motif**	
Positives:	39.0545 ± 36.3549	Positives:	119.2115 ± 23.1950
Negatives:	38.2500 ± 33.8585	Negatives:	72.6444 ± 83.7718

Distances from 2nd motif to 3rd motif		**Distances from 3rd motif to C-terminus**	
Positives:	157.9022 ± 23.8809	Positives:	29.4583 ± 14.5323
Negatives:	121.0278 ± 161.2368	Negatives:	37.7500 ± 29.5407

Based on this observation it is clear that the positive effect of distance penalty is due to the huge difference in SD between positives and negatives for these two distances. Noteworthy, in the case of GDSL lipases we have set expected distances of 50, 125, 177 and 50 residues for distances from the N-terminus to 1st motif, 1st to 2nd, 2nd to 3rd and 3rd to C-terminus respectively. As can be seen in Table A7 these distance were suboptimal since mean distances in the positives dataset are 39, 119, 158 and 30. Therefore, this has had a negative effect which is particularly strong in the case of the Strong **DP** which is very sensitive to small differences as illustrated in the case of the Desaturases First subfamily (2.3).

2.3. Desaturases First subfamily analysis

To further confirm the benefit of new parametrizations, we tested *Leitmotif's* performance on an additional dataset: the Desaturases First subfamily which is a member of the Fatty Acid (FA) Desaturases family. The FA Desaturases family (Pfam database PF00487; 21,042 sequences) comprises four distinct subfamilies (Desaturases First, Desaturases Omega, Desaturases Front-end, and Desaturases Sphingolipid; Hashimoto *et al*., 2008; Feng *et al*., 2017).

The Desaturases First subfamily has been selected for analysis since it has a previously described protein motif signature (Hashimoto *et al*., 2008), thus providing unbiased criteria for annotation. Additionally, for this subfamily we were able to retrieve the highest number (82/105) of previously annotated sequences described in the Hashimoto *et al*. study. The remaining missing sequences (23/105) were removed from the standardly used databases (UniProt, KEGG, NCBI) and have not been preserved in their archives (personal communication).

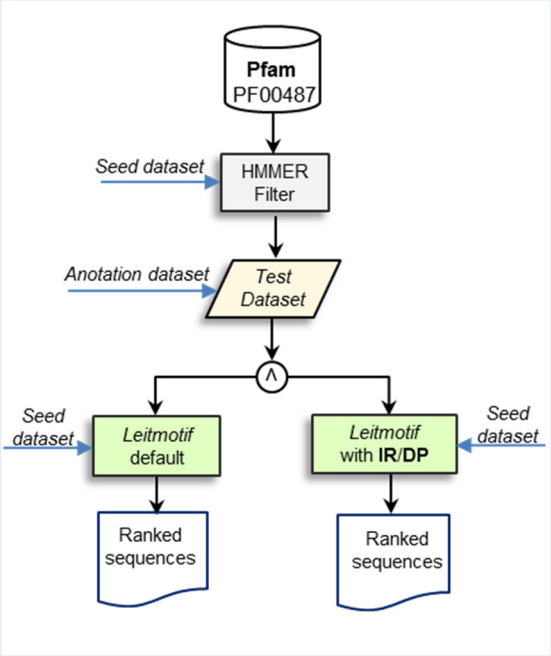
2.3.1. Datasets and experimental procedure

Lacking expert knowledge, and for purposes of unbiased annotation, firstly we divided the "Hashimoto dataset" (82 sequences) into two datasets:

(**i**) the *seed dataset* (**19**/82 sequences) reported as biochemically or genetically characterized in (Hashimoto *et al* 2008, Supplementary Data 1-1);

(**ii**) the *annotation dataset* (**63**/82 sequences) which was subsequently used to devise criteria for labelling sequences in the *test dataset* (described below) as positive or negative.

Fig. 5. Experimental procedure flowchart



First, in order to reduce the high number of easily discernible sequences belonging to the other three subfamilies, which would artificially inflate the ROC score (Gribskov and Robinson, 1996), HMMER (default settings with *seed dataset*) was used to filter them out from the Pfam PF00487 database (Fig. 5). All sequences which passed the HMMER default threshold were subsequently used as a *test dataset* (Supplement Desauturases First Test sequences). The *test dataset* (5509 sequences in total) was then divided into two subsets (4687 positives and 822 negatives) based on the annotation criteria (derived based on the *annotation set* only) and listed below. Next, the *seed dataset* was used for *Leitmotif* parameterization (both default as well as DP/IR parameterization) and *Leitmotif* was run both with selected **DP/IR** values and with the default parameterization (no **DP/IR**) (Fig. 5).

2.3.2. Annotation criteria

The annotation criteria used to label sequences in the *test dataset* as positive or negative was based on the *annotation dataset* (2.3.1). The annotation criteria are as follows:

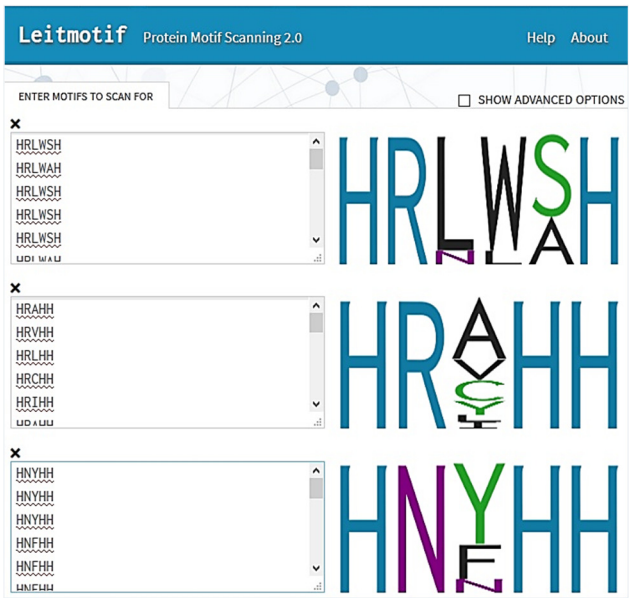
(**i**) *Motif signatures criterion*: a positive sequence has the motif signature defined by completely conserved motif residues in the *annotation dataset*.

Motif 1: **H** – (4X) – **H**
Motif 2: **H** – (2X) – **H** – **H**
Motif 3: (X) – **N** – (X) – **H** – **H**

(**ii**) *Motif distance criterion*: a positive sequence has a distance of **30 to 32** residues between its 1st and 2nd motif. This specific distance range was chosen since motif distances between the 1st and 2nd motif are strongly conserved (61/63 sequences have 31 residues between 1st and 2nd motif and 2/63 have 32 residues). The distance of 30 was also included since we were not able to retrieve the complete "Hashimoto dataset (23/105 were missing), as stated above.

2.3.3. Selection of DP/IR parameters

The seed motifs used with *Leitmotif* were extracted from the multiple sequence alignment (PROMALS) of the *seed dataset*. Figure 6 (left panel) shows sequence logos of seed motifs which have 11 completely conserved residues. Accordingly, all 11 of them were set as immutable (**IR**) for model parameterization. Note that only 8/11 residues are conserved in the *motif signature* (2.3.2). The right panel shows that distances in the seed between the 1st and 2nd motif are completely conserved (31 residues), distances from the 2nd to 3rd motif are somewhat conserved (127-136), whereas distances from 1st/3rd motif to N/C terminus are not conserved. Therefore, based on the seed the most appropriate choice for **DP** between the 1st and 2nd motif is Strong; for **DP** between the 2nd and 3rd motif is Medium or Weak, and None between 1st/3rd motif to N/C terminus. To isolate the effect of **DP** to a single parameter we used **DP** for the distance between the 1st and 2nd motif only.



93	HRLWSH	31	HRAHH	136	HNYHH	54
87	HRLWAH	31	HRCHH	136	HNYHH	63
113	HRLWSH	31	HRAHH	136	HNYHH	57
115	HRLWSH	31	HRAHH	136	HNYHH	57
118	HRLWSH	31	HRAHH	136	HNYHH	57
119	HRLWSH	31	HRAHH	136	HNYHH	57
119	HRLWSH	31	HRAHH	136	HNYHH	57
119	HRLWSH	31	HRAHH	136	HNYHH	57
96	HRLWSH	31	HRCHH	136	HNFHH	60
95	HRLWSH	31	HRCHH	136	HNFHH	60
85	HRLWSH	31	HRVHH	136	HNYHH	57
98	HRLWAH	31	HRVHH	136	HNYHH	80
119	HRLWAH	31	HRVHH	136	HNYHH	81
86	HRLWSH	31	HRLHH	135	HNYHH	62
83	HRLWAH	31	HRAHH	132	HNFHH	187
81	HRLWAH	31	HRAHH	132	HNFHH	195
160	HRLWSH	31	HRIHH	132	HNFHH	171
84	HRNLAH	31	HRYHH	127	HNNHH	49
82	HRNLAH	31	HRYHH	127	HNNHH	49

Fig. 6. **Seed motifs**. The left panel shows motif sequence logos generated by *Leitmotif* automatically after inserting motif alignments (19 sequences) into the motif input window. The right panel shows output results of the 19 seed sequences (*seed dataset*) scanned by *Leitmotif* (default parameterization) with distances between motif(s) and/or N/C-terminus. As depicted in the figure distances between the 1st and 2nd motifs are completely conserved (31 residues). Note that ranking scores are omitted from the figure due to limited space.

2.3.4. Results of Desaturases First subfamily analysis

Tables B1-B7 *Leitmotif's* performance on the *test dataset* with and without (IR/DP) parameters. As in the previous analysis, Relative (**R**) Sequence Weighting (**SW**) and *match* to *match* transition probability of 0.99 was used.

B1	ALG	IR	MD	DP	ROC	nROC 50	nROC 5	nROC 1
3 MOTIFS	MA	1 [H-R-(3X)-H], 2 [H-R-X-H-H], 3 [H-N-X-H-H]	103-31-134-79	/ - / - / - /	0.9271	0.5321	0.0536	0.0019
		[,,]		/ - / - / - /	0.8741	0.4959	0.0536	0.0019

B2	ALG	IR	MD	DP	ROC	nROC 50	nROC 5	nROC 1
3 MOTIFS	MA	[,,]	103-31-134-79	/ - S - / - /	0.9559	0.6110	0.4043	0.3643
		[,,]		/ - M - / - /	0.9328	0.5837	0.3873	0.3596
		[,,]		/ - W - / - /	0.8854	0.5098	0.1021	0.0060
		[,,]		/ - / - / - /	0.8741	0.4959	0.0536	0.0019

B3	ALG	IR	MD	DP	ROC	nROC 50	nROC 5	nROC 1
3 MOTIFS	MA	1 [H-R-(3X)-H], 2 [H-R-X-H-H], 3 [H-N-X-H-H]	103-31-134-79	/ - S - / - /	0.9885	0.9088	0.6419	0.5713
		1 [H-R-(3X)-H], 2 [H-R-X-H-H], 3 [H-N-X-H-H]		/ - M - / - /	0.9661	0.7608	0.4864	0.3652
		1 [H-R-(3X)-H], 2 [H-R-X-H-H], 3 [H-N-X-H-H]		/ - W- / - /	0.9306	0.5567	0.1039	0.0062
		[,,]		/ - / - / - /	0.8741	0.4959	0.0536	0.0019

B4	ALG	IR	MD	DP	ROC	nROC 50	nROC 5	nROC 1
2 MOTIFS	MA	1 [H-R-(3X)-H], 2 [H-R-X-H-H],	103-31-213	/ - / - / - /	0.7593	0.1023	0.0009	0
		[,]		/ - / - / - /	0.7555	0.1023	0.0009	0

B5	ALG	IR	MD	DP	ROC	nROC 50	nROC 5	nROC 1
2 MOTIFS	MA	[,]	103-31-213	/ - S - / - /	0.8707	0.1155	0.0010	0
		[,]		/ - M - / - /	0.8588	0.1155	0.0010	0
		[,]		/ - W - / - /	0.7624	0.1083	0.0010	0
		[,]		/ - / - / - /	0.7555	0.1023	0.0009	0

B6	ALG	IR	MD	DP	ROC	nROC 50	nROC 5	nROC 1
2 MOTIFS	MA	1 [H-R-(3X)-H], 2 [H-R-X-H-H],	103-31-213	/ - S - / - /	0.8723	0.1155	0.0010	0
		1 [H-R-(3X)-H], 2 [H-R-X-H-H],		/ - M - / - /	0.8618	0.1155	0.0010	0
		1 [H-R-(3X)-H], 2 [H-R-X-H-H],		/ - W - / - /	0.7655	0.1085	0.0010	0
		[,]		/ - / - / - /	0.7555	0.1023	0.0009	0

B7	ALG	IR	MD	DP	ROC	nROC 50	nROC 5	nROC 1
1 MOTIFS	MA	1 [H-R-(3X)-H]	103-244	/ - / - / - /	0.7409	0.0315	0.0003	0
		1 [ ]		/ - / - / - /	0.7389	0.0315	0.0003	0

Algorithm: **MA**-Modified Ancestral, **IR**-Immutable Residues; Description of motifs**: Motif 1** ([H-R-(3X)-H]; letters represent residues set as immutable (**IR**) according to seed motifs (Fig. 6 Left panel), His at position 1, Arg at position 2, and His at position 6); **Motif 2** ([H-R-X-H-H]; letters represent residues set as immutable (**IR**) according to seed motifs, His at position 1, Arg at position 2, and His at positions 5 & 6); **Motif 3** ([H-N-X-H-H]; letters represent residues set as immutable (**IR**) according to seed motifs, His at position 1, Asp at position 2, and His at positions 5 & 6). **MD**- Motif distances, **MD** were chosen based on seed sequences (Fig. 6 Right panel). **DP**-Distance Penalty strengths ("/"-None; "W"-Weak, "M"-Medium, "S"-Strong); Light grey rows show ROC scores with selected **IR/DP** values; Dark grey rows show scores without **DP** and **IR** parameters.

2.3.5. ROC curves with different parameterizations

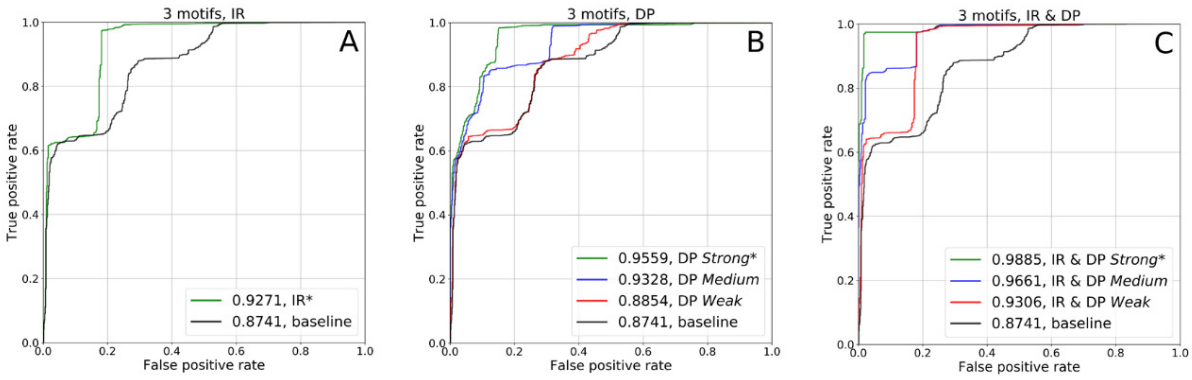


Fig. 7. (**A, B & C**): **Desaturases First subfamily ROC curves for 3 motifs**. Selected values for IR/DP are marked by asterisks. **A**) Results with Immutable Residues (**IR**) only. As shown in the figure setting **IR** substantially improves ROC score (i.e. area under the ROC curve)**. B**) Results with Distance Penalties (**DP**) only. As can be seen in the figure significant improvement in ranking is obtained by using **DP**. Moreover increasing the **DP** from None to Strong gradually improves the ROC scores (from 0.8741 to 0.9559 respectively)**. C**) Results of combining **DP** and **IR**. As shown in the figure ROC scores are substantially improved with the combination of **IR** and **DP**. Moreover, comparing Figures A, B, and C shows that the best overall ROC score is obtained when using the combination of Strong **DP** and **IR**.

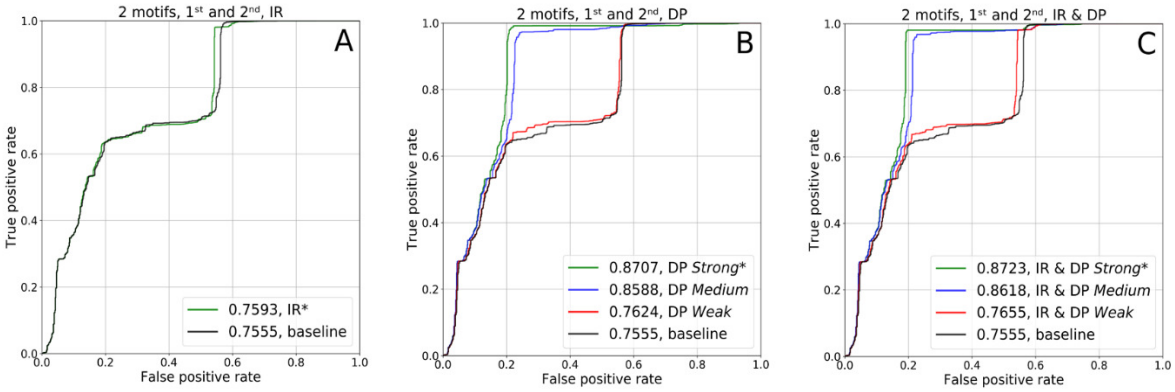


Fig. 8. (**A, B & C**): **Desaturases First subfamily ROC curves for two motifs** (**1st & 2nd**). Selected values for IR/DP are marked by asterisks. **A**) Results with Immutable Residues (**IR**) only. As can be seen in this setting the difference between ROC scores is very modest (0.7555 vs 0.7593). **B**) Results for Distance Penalties (**DP**) only. As can be seen significant improvement in ranking is obtained by using **DP**. Increasing the **DP** from None to Strong increases ROC scores. **C**) Results with the combination of **DP** and **IR**. Again it is clearly visible that the ROC score is improved by using **DP** and **IR** vs no **IR/DP**. In line with the results shown in A, the additional benefit of **IR** is very modest.

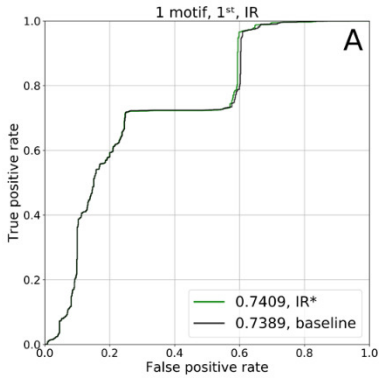


Fig. 9. **Desaturases First subfamily ROC curves for one motif** (**1st**). Selected values for IR/DP are marked by asterisks. Results with only Immutable Residues (**IR**) and 1st motif. As shown in the figure, setting residues as immutable improves the ROC score very slightly.

In summary, using **IR/DP** increases ROC scores in all cases (Fig. 7, 8 & 9). As expected, best results are obtained by using all three motifs as in the GDSL case. Analogously, we have included additional cases with one and two motifs. This was done in order to illustrate that **DP** & **IR** can improve ROC scores regardless of the number of motifs. Expectedly, discarding motifs limits the benefit of **IR** as can be seen by comparing Figures 7A, 8A and 9.

Next, we analysed the unusual stepwise shape of ROC curves (Fig. 7, 8 & 9). This is due to the fact that there is a large number of negative sequences in the *test dataset* which have the Desaturases First motif signature with a distance of 34 residues between the 1st and 2nd motif. These sequences were annotated as negatives since they don't satisfy the motif distance criterion (2.3.2). Prompted by this we analysed all distances between the 1st and 2nd motif in the *test dataset*. Their histogram is shown in Figure 10. As can be seen, there is a very strong peak at 31 (which was expected taking into account the Hashimoto *annotation dataset* as well as the *seed dataset*). However, the strong peak around 34 is very unusual since the number of sequences with a distance of 33 is negligible (taking into account the size of the dataset). In light of this, and because the Hashimoto *annotation dataset* as well as *seed dataset* comprise evolutionary divergent sequences, it seems plausible that the sequences with a distance of 34 comprise another as yet undescribed subfamily within the Desaturases First subfamily. However, to investigate this hypothesis further is beyond the scope of this manuscript.

Note also that there is a huge number of sequences whose distances between the 1st and 2nd motif range from 30 to 32 (4998 in total), whereas the number of sequences with distances beyond this range (apart from the 34 case) drops sharply. This strongly indicates that the 30 to 32 motif distance criterion, defined based on the Hashimoto *annotation dataset,* is indeed appropriate.

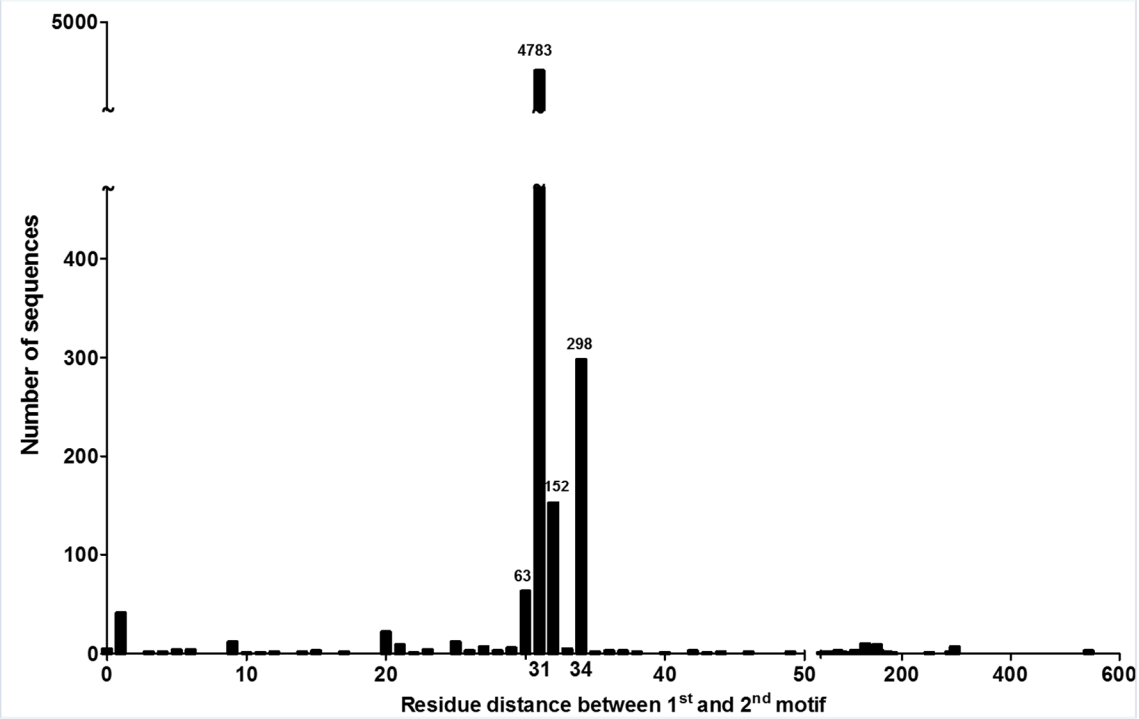


Fig. 10. Distances between the 1st and 2nd motif in the *test dataset*.

References:

Durbin,R. *et al.* (1998) Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, Cambridge University Press, Cambridge, UK.

Gribskov,M. and Robinson, N.L. (1996) Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. Comput. Chem., 20, 25–33.

Grundy,W.N. et al., (1997) meta-MEME: Motif-based hidden Markov models of protein families. Comput Appl Biosci. 13(4):397–406.

Feng,J. *et al.* (2017) Genome-wide identification of membrane-bound fatty acid desaturase genes in *Gossypium hirsutum* and their expressions during abiotic stress. Sci Rep 7, 45711.

Hashimoto,K. *et al.* (2008) The repertoire of desaturases and elongases reveals fatty acid variations in 56 eukaryotic genomes. J Lipid Res., 49(1):183-91.

Henikoff,J.G. and Henikoff, S. (1996) Using substitution probabilities to improve position-specific scoring matrices. *Comput. Appl. Biosci*., **12**(2), 135-43.

Leščić Ašler,I. et al. (2017) Catalytic Dyad in the SGNH Hydrolase Superfamily: In-depth Insight into Structural Parameters Tuning the Catalytic Process of Extracellular Lipase from Streptomyces rimosus. ACS Chem. Biol., 12, 1928–1936

Sjölander, K. *et al.* (1996) Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Comput. Appl. Biosci.*, **12**(4), 327-45.

Upton,C. and. Buckley,J.T. (1995) A new family of lipolytic enzymes? Trends. Biochem. Sci., 20 178 e179

Vujaklija,I. *et al.* (2016) An effective approach for annotation of protein families with low sequence similarity and conserved motifs: identifying GDSL hydrolases across the plant kingdom. *BMC Bioinformatics,* **18,** 17–91.