A Biological Background

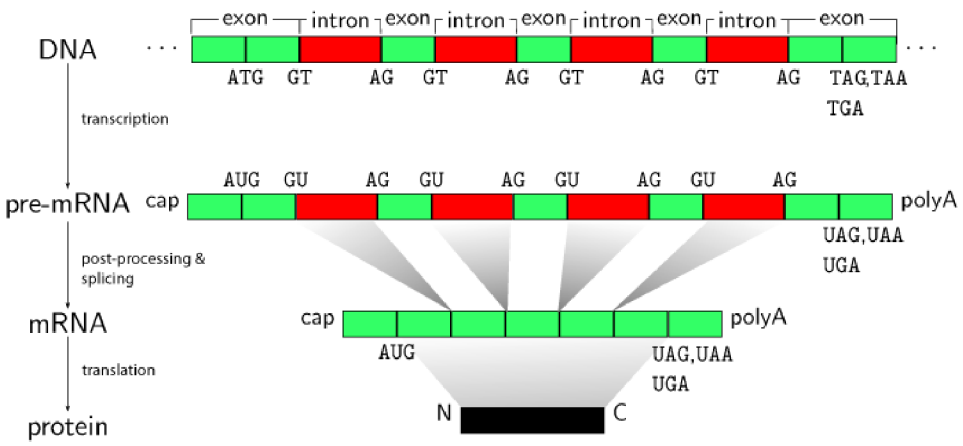


Figure A1: The major steps in protein synthesis: transcription, post-processing and translation. In the post-processing step, the pre-mRNA is transformed into mRNA. One necessary step in the process of obtaining mature mRNA is called *splicing*. The mRNA sequence of a eukaryotic gene is "interrupted" by non-coding regions called *introns*. A gene starts with an exon and may then be interrupted by an intron, followed by another exon, intron and so on until it ends in an exon. In the splicing process, introns are removed. As a result, there are two different splice sites: the exon-intron boundary, referred to as the donor site or 5' site (of the intron) and the intron-exon boundary, that is the acceptor or 3' site. Splice sites have quite strong consensus sequences, i.e. almost each position in a small window around the splice site is representative of the most frequently occurring nucleotide when many existing sequences are compared in an alignment. For example, the 5' site's consensus is $\mathtt{A}_{64}\mathtt{G}_{73}\mathtt{G}_{100}\mathtt{T}_{100}\mathtt{A}_{62}\mathtt{A}_{68}\mathtt{G}_{84}\mathtt{T}_{63}$, while the 3' site's consensus is $\mathtt{C}_{65}\mathtt{A}_{100}\mathtt{G}_{100}$, where the subscripts denote the frequency of the symbol in percent. The dimers GT and AG can therefore be used to identify potential donor and acceptor sites.

B How Different Are the Problems?

To better understand the distance between the organisms with respect to the specific task of splice site detection, we performed experiments to classify the example sequences according to their origin. For these pairwise discrimination tasks, we used the almost same method as for the splice site detection task, namely SVMs with the weighted degree kernel ($\ell = 22$). As a performance measure we used the area under the precision-recall curve (auPRC). In this context, a higher auPRC is associated with a greater difference between the functioning of the splice mechanism. We examined the difference between true splice sites, as well as between decoy sites (i.e. other genomic sequences). For each classifier we used $1,000$ examples from each of the two investigated organisms for training (67%) and evaluation (33%). The results are shown in Table A1. We observe that the discrimination performance of discriminating true splice sites from different organisms correlates well with their evolutionary distance. The discrimination performance for decoy examples saturates for organisms more than 100-200 million years apart (the rest of the genome is evolving faster than important functional elements such as splice sites).

	time (10^6 years)	**auPRC+**	**auPRC-**
C. elegans - *C. remanei*	> 100	63%	67%
C. elegans - *P. pacificus*	200	93%	81%
C. elegans - *D. melanogaster*	990	95%	78%
C. elegans - *A. thaliana*	$> 1,100$	98%	77%

Table A1: Summary of different distance measures between source and target organism. Reported are the time since divergence. Additionally the performance of classifiers that distinguishes between examples of two organisms. The performance measure is give as auPRC+ for a classifier that distinguished between true splice sites and auPRC- for a classifier that distinguished between decoy sites

C Hyper-parameter Selection

Table A2 shows, which hyper-parameters have been optimized for each method. Hyper-parameter tuning was performed using a grid search.

		hyper-parameter	total
SVM_S	C_S	0.1, 0.2, 0.4, 0.7, 1.4, 2.7, 5, 10, 19, 37, 72, 139, 268, 518, 1000	15
SVM_T	C_t	0.1, 0.2, 0.4, 0.7, 1.4, 2.7, 5, 10, 19, 37, 72, 139, 268, 518, 1000	15
SVM_S+SVM_T	C_S	0.1, 0.2, 0.4, 0.7, 1.4, 2.7, 5, 10, 19, 37, 72, 139, 268, 518, 1000	
	C_t	0.1, 0.2, 0.4, 0.7, 1.4, 2.7, 5, 10, 19, 37, 72, 139, 268, 518, 1000	$15\times15\times21$
	α	$[0 : 0.05 : 1.0]$	$= 4725$
SVM_{S+T}	C_S	0.001, 0.01, 0.1, 0.8, 7, 100, 1000	7×7
	C_t	0.001, 0.01, 0.1, 0.8, 7, 100, 1000	$= 49$
$\text{SVM}_{S,T}$	C_S	0.001, 0.01, 0.1, 0.8, 7, 100, 1000	7
$\text{SVM}_{S\to T}$	C	0.001, 0.01, 0.1, 0.8, 7, 100, 1000	7×5
	α	0, 0.25, 0.5, 0.75, 1	$= 35$
$\text{SVM}_{S\times T}$	C	0.01, 0.1, 0.3, 0.8, 2.4, 7, 20	7×8
	B	0.1, 0.3, 0.8, 2.4, 7, 20, 50, 100	$= 56$
M-SVM_S+SVM_T	C_S	0.1, 0.2, 0.4, 0.7, 1.4, 2.7, 5, 10, 19, 37, 72, 139, 268, 518, 1000	
	C_t	0.1, 0.2, 0.4, 0.7, 1.4, 2.7, 5, 10, 19, 37, 72, 139, 268, 518, 1000	$15\times15\times21$
	α	$[0 : 0.05 : 1.0]$	$= 4725$
M-$\text{SVM}_{S,T}$	C_S	0.001, 0.01, 0.1, 0.8, 7, 100, 1000	7×7
	C_t	0.001, 0.01, 0.1, 0.8, 7, 100, 1000	$= 49$
M-$\text{SVM}_{S\to T}$	C	0.001, 0.01, 0.1, 0.8, 7, 100, 1000	7×5
	α	0, 0.25, 0.5, 0.75, 1	$= 35$

Table A2: For each of the proposed method we performed a grid search for the respective hyper-parameters. The search space for the appropriate hyper-parameters is given. All hyper-parameters are tuned on an evaluation set (33% of available target data).

D Tabular Single-Source Results

	2500	6500	16000	40000	100000
$SVM_{S,T}$	**77.06 (±2.13)**	**77.80 (±2.89)**	**77.89 (±0.29)**	**79.02 (±0.09)**	**80.49 (±0.0)**
SVM_S+SVM_T	71.61 (±2.39)	76.58 (±0.52)	77.50 (±0.77)	78.19 (±0.44)	80.19 (±0.0)
SVM_{SxT}	75.37 (±2.56)	76.10 (±2.62)	76.76 (±0.28)	77.82 (±0.23)	79.75 (±0.0)
$SVM_{S\rightarrow T}$	75.78 (±8.43)	75.55 (±1.08)	77.23 (±0.37)	78.11 (±0.49)	79.84 (±0.0)
SVM_{S+T}	75.49 (±1.88)	75.87 (±0.25)	77.23 (±0.47)	77.33 (±0.83)	79.96 (±0.0)
SVM_S	75.52 (±0.44)	76.27 (±0.42)	75.65 (±0.23)	75.65 (±0.4)	75.65 (±0.0)
SVM_T	24.04 (±4.53)	46.45 (±2.29)	60.51 (±1.01)	70.50 (±0.85)	78.04 (±0.0)

Table A3: *C. remanei*

	2500	6500	16000	40000	100000
$SVM_{S,T}$	64.72 (±3.75)	66.39 (±0.66)	**68.44 (±0.67)**	71.00 (±0.38)	74.88 (±0.0)
SVM_S+SVM_T	**64.74 (±3.49)**	**67.30 (±1.38)**	66.58 (±1.82)	**71.82 (±0.97)**	**75.39 (±0.0)**
SVM_{SxT}	57.67 (±26.14)	66.33 (±0.28)	67.29 (±2.24)	71.46 (±0.21)	74.99 (±0.0)
$SVM_{S\rightarrow T}$	63.52 (±14.05)	66.07 (±0.07)	67.59 (±1.7)	70.90 (±0.95)	75.11 (±0.0)
SVM_{S+T}	62.99 (±1.48)	65.87 (±0.83)	68.02 (±0.97)	70.85 (±0.37)	74.73 (±0.0)
SVM_S	62.73 (±0.62)	63.77 (±0.04)	63.75 (±0.75)	63.77 (±0.36)	63.83 (±0.0)
SVM_T	20.36 (±3.94)	38.16 (±3.21)	57.28 (±3.46)	67.90 (±1.35)	74.10 (±0.0)

Table A4: *P. pacificus*

	2500	6500	16000	40000	100000
$SVM_{S,T}$	**40.80 (±2.18)**	37.87 (±3.77)	**52.33 (±0.91)**	**58.17 (±1.5)**	63.26 (±0.0)
SVM_S+SVM_T	37.23 (±1.58)	40.36 (±3.32)	48.64 (±0.99)	54.38 (±1.57)	62.26 (±0.0)
SVM_{SxT}	38.71 (±7.67)	**41.23 (±1.4)**	49.58 (±0.91)	56.20 (±1.86)	62.22 (±0.0)
$SVM_{S\rightarrow T}$	35.29 (±6.72)	40.15 (±2.47)	48.98 (±2.19)	54.60 (±1.99)	**63.53 (±0.0)**
SVM_{S+T}	36.43 (±1.18)	37.98 (±4.05)	49.46 (±1.38)	56.56 (±2.36)	62.07 (±0.0)
SVM_S	32.95 (±0.38)	33.05 (±0.07)	33.07 (±0.25)	33.07 (±0.01)	33.74 (±0.0)
SVM_T	14.59 (±1.02)	26.69 (±0.58)	38.33 (±2.06)	51.32 (±2.86)	61.26 (±0.0)

Table A5: *D. melanogaster*

	2500	6500	16000	40000	100000
$SVM_{S,T}$	24.21 (±3.41)	27.30 (±1.46)	**38.49 (±1.59)**	**49.75 (±1.46)**	**56.54 (±0.0)**
SVM_S+SVM_T	21.70 (±2.77)	**28.55 (±1.96)**	35.80 (±1.48)	44.07 (±2.99)	54.06 (±0.0)
SVM_{SxT}	**24.62 (±3.07)**	27.33 (±3.17)	38.20 (±1.32)	47.05 (±2.39)	53.60 (±0.0)
$SVM_{S\rightarrow T}$	17.09 (±6.79)	26.41 (±4.81)	36.83 (±1.74)	47.98 (±2.25)	55.99 (±0.0)
SVM_{S+T}	20.06 (±3.23)	24.71 (±3.25)	37.72 (±1.74)	47.31 (±2.55)	53.41 (±0.0)
SVM_S	14.07 (±0.46)	14.85 (±0.1)	14.23 (±0.53)	14.83 (±0.49)	14.33 (±0.0)
SVM_T	10.23 (±1.56)	19.07 (±2.53)	32.56 (±1.91)	45.34 (±2.83)	53.63 (±0.0)

Table A6: *A. thaliana*

E Tabular Multi-Source Results

	2500	6500	16000	40000	100000
M-SVM$_{S,T}$	**69.45 (±0.17)**	71.44 (±1.5)	71.03 (±1.8)	**76.21 (±0.2)**	79.11 (±0.0)
M-SVM$_S$+SVM$_T$	68.51 (±2.95)	**72.69 (±0.86)**	**72.78 (±0.8)**	75.75 (±0.64)	79.06 (±0.0)
M-SVM$_{S\rightarrow T}$	63.23 (±5.61)	70.11 (±0.52)	72.09 (±0.19)	75.32 (±0.32)	**79.24 (±0.0)**
SVM$_T$	24.04 (±4.53)	46.45 (±2.29)	60.51 (±1.01)	70.50 (±0.85)	78.04 (±0.0)

Table A7: *C. remanei*

	2500	6500	16000	40000	100000
M-SVM$_{S,T}$	61.38 (±2.05)	64.07 (±1.07)	**67.81 (±2.0)**	**71.05 (±1.26)**	**75.15 (±0.0)**
M-SVM$_S$+SVM$_T$	**62.88 (±0.3)**	**64.77 (±0.52)**	65.52 (±0.81)	70.50 (±0.94)	74.20 (±0.0)
M-SVM$_{S\rightarrow T}$	61.46 (±0.49)	62.48 (±0.41)	64.78 (±1.7)	70.14 (±0.56)	74.43 (±0.0)
SVM$_T$	20.36 (±3.94)	38.16 (±3.21)	57.28 (±3.46)	67.90 (±1.35)	74.10 (±0.0)

Table A8: *P. pacificus*

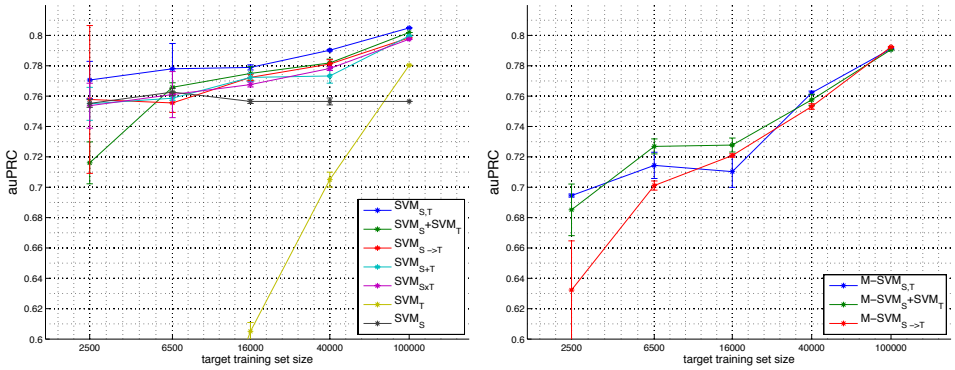
	2500	6500	16000	40000	100000
M-SVM$_{S,T}$	46.32 (±0.39)	47.71 (±1.03)	53.17 (±0.45)	**57.56 (±1.54)**	**62.66 (±0.0)**
M-SVM$_S$+SVM$_T$	**46.61 (±3.27)**	**48.15 (±3.02)**	52.12 (±0.73)	57.01 (±1.64)	62.12 (±0.0)
M-SVM$_{S\rightarrow T}$	40.89 (±2.28)	44.61 (±1.51)	**53.29 (±1.47)**	56.35 (±1.57)	61.57 (±0.0)
SVM$_T$	14.59 (±1.02)	26.69 (±0.58)	38.33 (±2.06)	51.32 (±2.86)	61.26 (±0.0)

Table A9: *D. melanogaster*

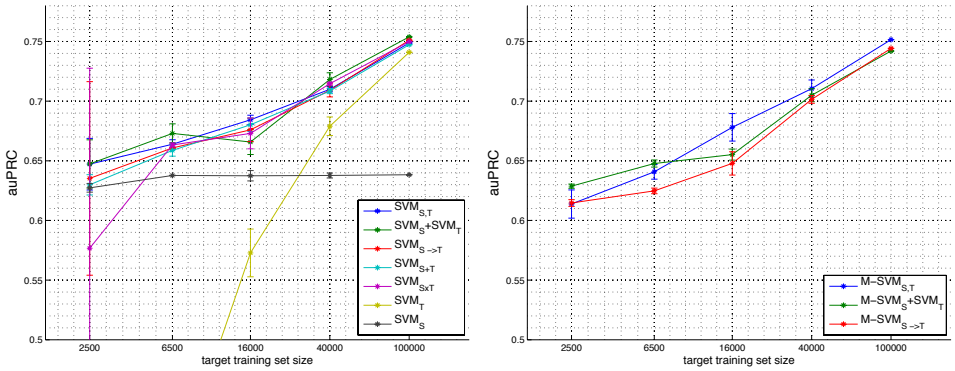
	2500	6500	16000	40000	100000
M-SVM$_{S,T}$	**30.90 (±2.12)**	**36.43 (±3.46)**	**43.63 (±1.92)**	**49.49 (±1.92)**	**56.57 (±0.0)**
M-SVM$_S$+SVM$_T$	26.61 (±2.29)	35.58 (±0.31)	39.43 (±1.98)	46.98 (±3.73)	54.11 (±0.0)
M-SVM$_{S\rightarrow T}$	27.17 (±1.33)	33.18 (±3.32)	39.32 (±2.07)	47.53 (±2.2)	55.50 (±0.0)
SVM$_T$	10.23 (±1.56)	19.07 (±2.53)	32.56 (±1.91)	45.34 (±2.83)	53.63 (±0.0)

Table A10: *A. thaliana*

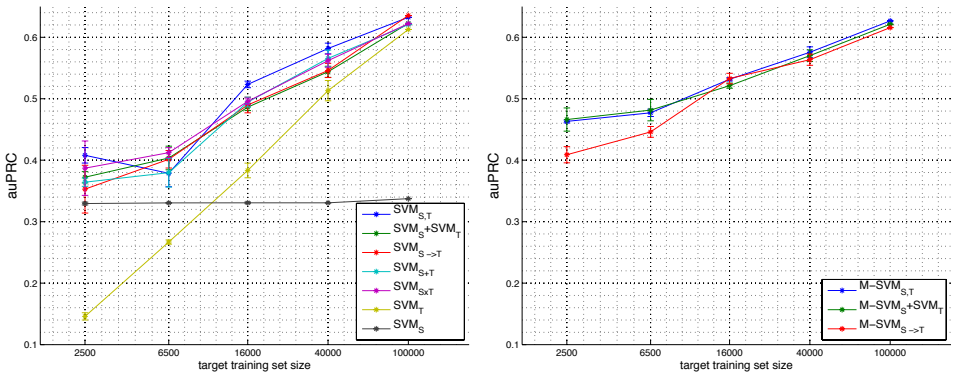
C. remanei (closely related to *C. elegans*)



P. pacificus (relatively closely related to *C. elegans*)



D. melanogaster (distantly related to *C. elegans*)



A. thaliana (most distantly related to *C. elegans*)

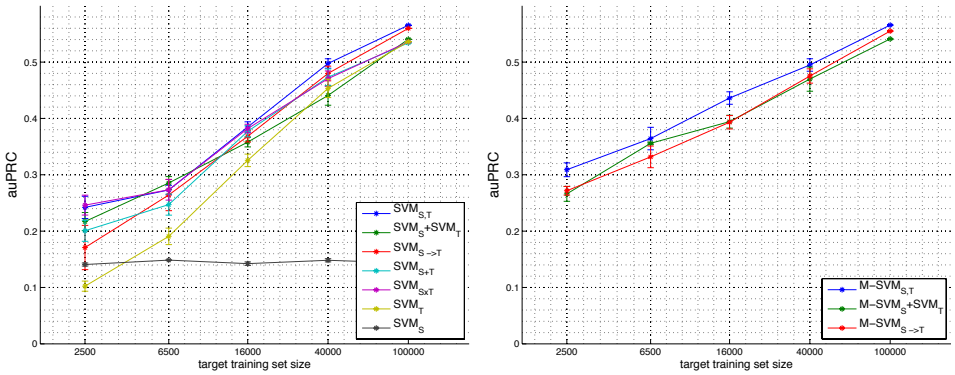


Figure A2: Area under the precision-recall curve for varying target training set sizes for the four considered target organisms. Shown are the median values over up to three different training set splits and the error bars indicate the confidence intervals.