1 Column Generation Method for Structured Output Learning

Algorithm 1 Column Generation Method for Structured Output Learning

1: $\mathbf{w}^{(1)} = \mathbf{w}_p$
2: $k = 1$ and $\Gamma_i = \emptyset \quad \forall i$
3: **repeat**
4: **for** $i = 1, .., n$ **do**
5: $\boldsymbol{y}^* = \text{argmax}_{\boldsymbol{y} \in \Upsilon} \{\langle \mathbf{w}^{(k)}, \Psi(\boldsymbol{x}_i, \boldsymbol{y}) \rangle + \Delta(\boldsymbol{y}_i, \boldsymbol{y})\}$
6: **if** $\langle \mathbf{w}^{(k)}, \Psi(\boldsymbol{x}_i, \boldsymbol{y}^*) \rangle + \Delta(\boldsymbol{y}_i, \boldsymbol{y}^*) > \max_{(\Psi, \Delta) \in \Gamma_i} \{\langle \mathbf{w}^{(k)}, \Psi \rangle + \Delta\}$ **then**
7: $\Gamma_i = \Gamma_i \cup (\Psi(\boldsymbol{x}_i, \boldsymbol{y}^*), \Delta(\boldsymbol{y}_i, \boldsymbol{y}^*))$
8: **end if**
9: **end for**
10: $\mathbf{w}^{(k)} = \underset{\mathbf{w} \in \mathcal{H}}{\text{argmin}} \quad \left\{ R_{p,\gamma}(\mathbf{w}) + C \sum_{i=1}^{n} \ell \left(\max_{(\Psi, \Delta) \in \Gamma_i} \{\langle \mathbf{w}, \Psi \rangle + \Delta\} - \langle \mathbf{w}, \Psi(\boldsymbol{x}_i, y_i) \rangle \right) \right\}$
11: $k = k + 1$
12: **until** no changes in $(\Psi_i, \Delta_i)\ \forall i$

2 Details on Eukaryotic Gene Finding

These sections provide additional detail on our method for inferring the exon-intron structure of eukaryotic genes using RNA-seq data and computational splice site predictions.

RNA-seq data for *C. elegans* was aquired using a strand-specific, paired-end protocol for Illumina sequencing (2x76bp read length).

Subsequently, RNA-seq reads were mapped to the genome using methods able to align reads across splice junctions (we used [3], other possibilities include [2, 7]). Moreover, we exploited splice site predictions made from the genome sequence with SVMs and string kernels as described previously [6]. Their incorporation into gene prediction was done in a way that is very similar to a recently developed gene finding system [5].

The structured-output inference methods used here extend a previously published approach [9] based on hidden Markov support vector machines [1, 8]. In essence, feature values from the input sequence $\boldsymbol{x}$ are transformed by piecewise linear functions, which depend on the state $\boldsymbol{y}_i$ and feature type. These thus make up an integral part of $\Psi(\boldsymbol{x}, \boldsymbol{y})$, and their parametrization constitutes part of the parameter vector $\mathbf{w}$ [9, 4]. The result of these feature transformations is linearly combined in a discriminant function $F(\boldsymbol{x}, \boldsymbol{y}; \mathbf{w})$ to yield $F(\boldsymbol{x}, \boldsymbol{y}; \mathbf{w}) = \langle \mathbf{w}, \Psi(\boldsymbol{x}, \boldsymbol{y}) \rangle$ [1, 8, 9]. Since the discriminant function assigns a real-valued score to a combination of feature sequence $\boldsymbol{x}$ and label sequence

$\boldsymbol{y}$, decoding for its argmax yields the highest-scoring label sequence with for a given $\boldsymbol{x}$ and $\mathbf{w}$ [1, 8, 9].

2.1 Features

The following features derived from RNA-seq read alignments (independent of the label sequence $\boldsymbol{y}$) were used as input sequence $\boldsymbol{x}$:

- number of reads aligned at the given position, indicating an exon.
- number of spliced reads that span the given position (strand-specific), indicating an intron.
- number of spliced reads supporting a donor splice site at the given position (strand-specific).
- number of spliced reads supporting an acceptor splice site (strand-specific).
- number of paired-read alignments spanning the given position (if read pair information is available, strand-specific), an indicator of transcript connectivity.

Moreover, we used splice site prediction features (see [6] for details):

- donor splice site prediction transformed to a probabilistic confidence (strand-specific).
- probabilistic acceptor splice site prediction confidence (strand-specific).

2.2 State model

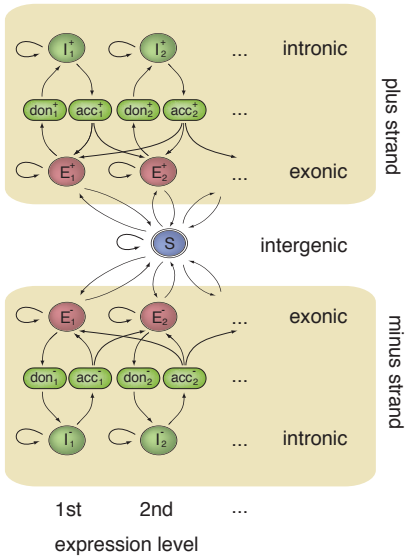

Figure 1: State model utilized for eukaryotic gene finding. Ovals correspond to states and arrows indicate allowed transitions. The correspondence between states and atomic labels is color-coded. The first and last intron states are associated with splice site signals at exon-intron junctions. The model is strand-specific and consists of econfirmation-specific submodels (columns and state-subscripts, see label at bottom) which allows optimizing different parameter sets depending on the experimental support of a given gene.

2.3 Loss function

The loss function captures our problem-specific knowledge about how much a given label sequence deviates and from the correct one and how the margin should be rescaled. The loss $\Delta(\boldsymbol{y}, \hat{\boldsymbol{y}})$ between the correct label sequence $\boldsymbol{y}$ and any predicted labeling $\hat{\boldsymbol{y}}$ is composed of position-wise loss terms:

$$\Delta(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \sum_{i=1}^{|\boldsymbol{y}|} \delta_b(y_i, \hat{y}_i) + \delta_s(y_i, \hat{y}_i) + \delta_{ei}(y_i, \hat{y}_i) + \delta_{fp}(y_i, \hat{y}_i) + \delta_{fn}(y_i, \hat{y}_i).$$

Most importantly, we penalize wrong segment boundaries by an intron-boundary loss:

$$\delta_b(y_i, \hat{y}_i) = \begin{cases} 10 & \text{if } y_i \neq \hat{y}_i \text{ and any of them is the first or last intron position} \\ & \text{(any } \texttt{acc} \text{ or } \texttt{don} \text{ state, see Fig. 1),} \\ 0 & \text{otherwise.} \end{cases}$$

Additionally there is a loss for predictions on the wrong strand:

$$\delta_s(y_i, \hat{y}_i) = \begin{cases} 0.1 & \text{if } y_i \text{ and } \hat{y}_i \text{ correspond to different strands (see Fig. 1),} \\ 0 & \text{otherwise,} \end{cases}$$

a loss for exon-intron confusions:

$$\delta_{ei}(y_i, \hat{y}_i) = \begin{cases} 0.1 & \text{if either } y_i \text{ or } \hat{y}_i \text{ is an exon state} \\ & \text{and the other one an intron state (see Fig. 1),} \\ 0 & \text{otherwise,} \end{cases}$$

a loss for positions in false-positive gene predictions:

$$\delta_{fp}(y_i, \hat{y}_i) = \begin{cases} 0.5 & \text{if } y_i \neq \hat{y}_i \text{ and } y_i = S \text{ (see Fig. 1)} \\ 0 & \text{otherwise,} \end{cases}$$

and a loss for positions in false-negative gene predictions:

$$\delta_{fn}(y_i, \hat{y}_i) = \begin{cases} 0.5 & \text{if } y_i \neq \hat{y}_i \text{ and } \hat{y}_i = S \text{ (see Fig. 1)} \\ 0 & \text{otherwise.} \end{cases}$$

Furthermore there are smaller loss terms (< 0.05, omitted above) for exon states with incorrect expression level.

3 Tabular Results for Prokaryotic MTL

	Independent	(std)	Union	(std)	MTL	(std)
E.coli	0.7985	0.0445	0.9501	0.0100	0.9573	0.0125
E.fergusonii	0.7262	0.0303	0.9573	0.0065	0.9651	0.0027
A.tumefaciens	0.7796	0.0744	0.8707	0.0362	0.9112	0.0111
H.pylori	0.7760	0.0983	0.9424	0.0087	0.9533	0.0133
B.anthracis	0.8387	0.0490	0.9184	0.0257	0.9234	0.0198
B.subtilis	0.7652	0.1751	0.9399	0.0243	0.9453	0.0283
M.smithii	0.9048	0.0136	0.9392	0.0106	0.9416	0.0072
S.islandicus	0.8151	0.0595	0.9149	0.0115	0.9263	0.0096
mean	0.8005	0.0681	0.9291	0.0167	0.9404	0.0131

References

[1] Y. Altun, I. Tsochantaridis, and T. Hofmann. Hidden Markov support vector machines. *Proceedings of ICML*, 2003.

[2] F. D. Bona, S. Ossowski, K. Schneeberger, and G. Rätsch. Optimal spliced alignments of short sequence reads. *Bioinformatics*, 24(16):i174–80, 2008.

[3] G. Jean, A. Kahles, V. T. Sreedharan, F. D. Bona, and G. Rätsch. RNA-seq read alignments with PALMapper. *Current Protocols in Bioinformatics*, 32:11.6.1–11.6.38, 2010.

[4] G. Rätsch and S. Sonnenburg. Large scale hidden semi-Markov SVMs. *Proceedings of NIPS*, 2007.

[5] G. Schweikert, A. Zien, G. Zeller, J. Behr, C. Dieterich, C. S. Ong, P. Philips, F. D. Bona, L. Hartmann, A. Bohlen, N. Krüger, S. Sonnenburg, and G. Rätsch. mGene: Accurate SVM-based gene finding with an application to nematode genomes. *Genome Research*, 2009.

[6] S. Sonnenburg, G. Schweikert, P. Philips, J. Behr, and G. Rätsch. Accurate splice site prediction using support vector machines. *BMC Bioinformatics*, 8 Suppl 10:S7, 2007.

[7] C. Trapnell, L. Pachter, and S. L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[8] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 2006.

[9] G. Zeller, S. R. Henz, S. Laubinger, D. Weigel, and G. Rätsch. Transcript normalization and segmentation of tiling array data. *Pacific Symposium on Biocomputing*, pages 527–38, 2008.