Bioinformatics

(Past, Current and Future Projects)

Sören Sonnenburg

Friedrich Miescher Laboratory, Tübingen

joint work with

Alexander Zien, Jonas Behr, Gabriele Schweikert, Vojtech Franc, Petra Philips and Gunnar Rätsch

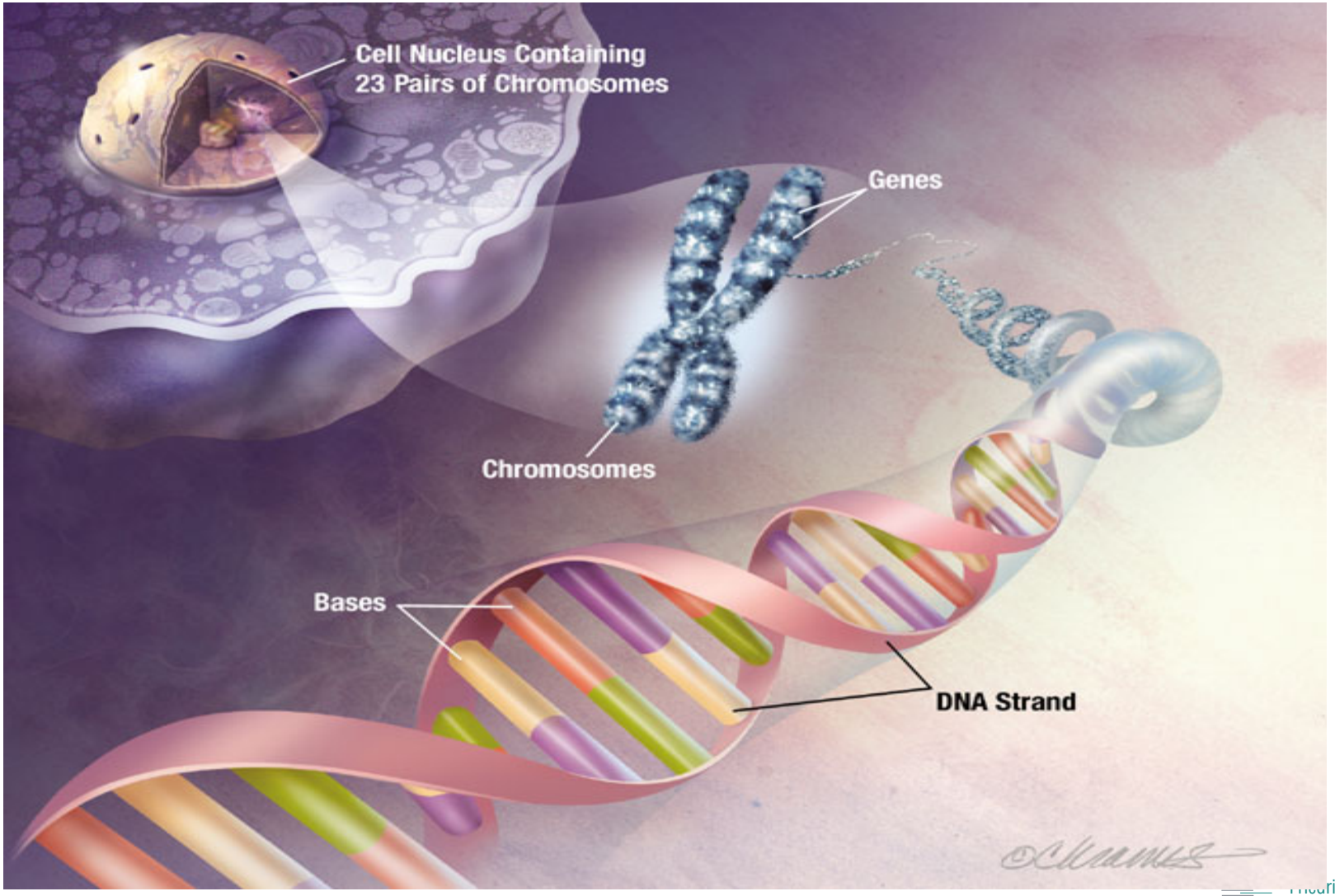
Friedrich Miescher Laboratory

of the Max Planck Society

Outline

1. Goal
2. State of Affairs
3. Todo

Unravel Genome, Location and Regulation of Genes I



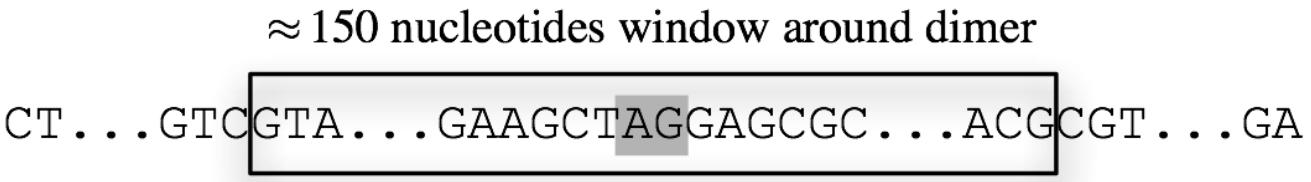
Unravel Genome, Location and Regulation of Genes II

1. Localize Genomic Signals (Start, End, Coding Parts)
2. Content prediction (Is this sequence coding or non-coding?)
3. Translate knowledge to other Genomes (DA $\rightarrow$ Chris)
4. Combine Signal/Content Prediction to predict Gene Structures
5. Interpret Learned Results

$\Rightarrow$ **Submitted DFG Proposal with Alex, Gunnar, Klaus** ...

Localize Genomic Signals *faaast*

Discriminate true signal positions against all other positions

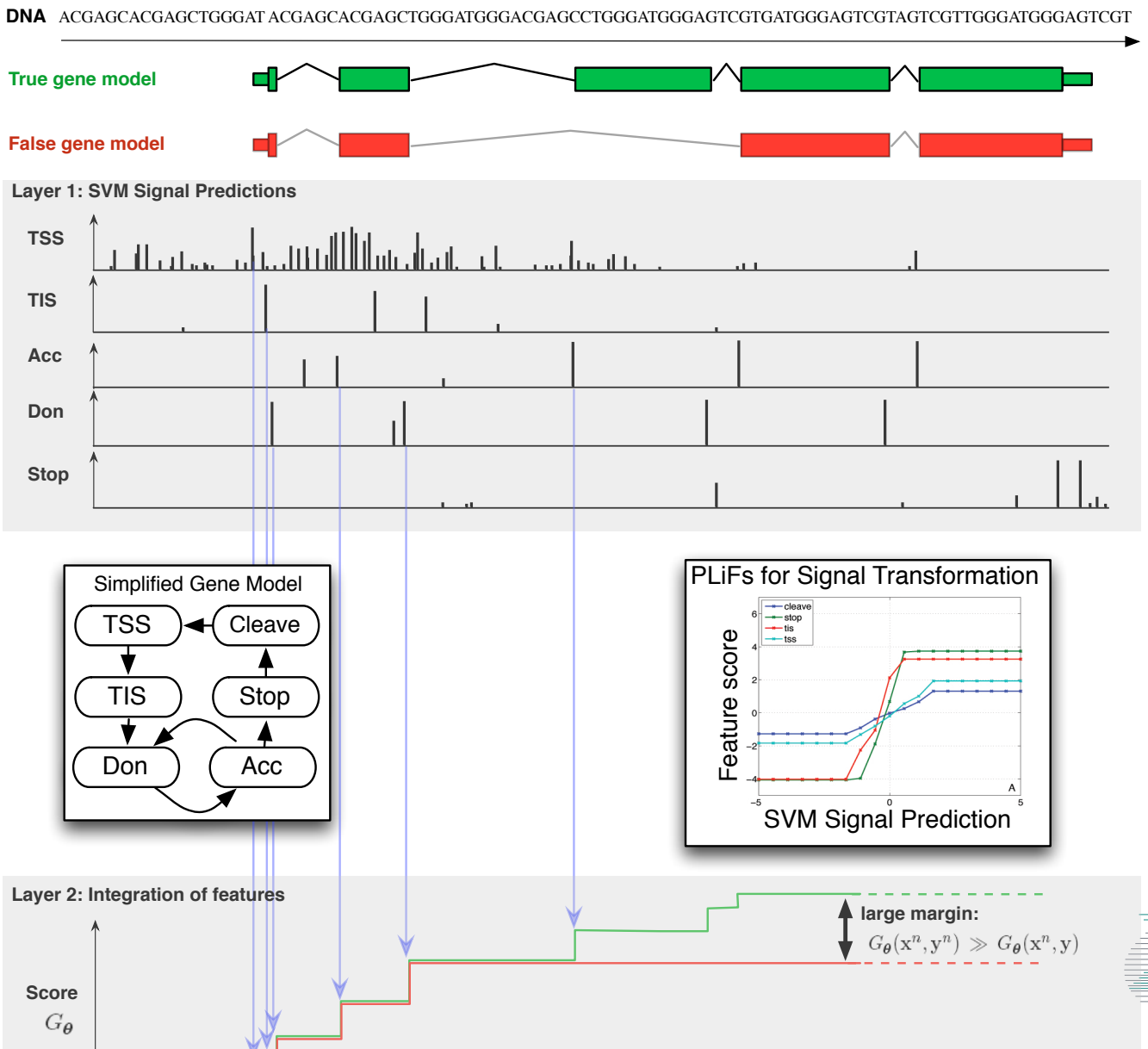


State of Signal/Content Detection

- SVM+String kernels + `linadd` acceleration works OK
- Recent work: Further drastic speedup using SVM-Ocas in primal
  - Idea: Train SVM in primal using kernel feature space (1 WD kernel, 2 Weighted Spectrum Kernels)
  - Problem: $>$ 12 million dims; 50 million examples
  - Only $\mathbf{w} \leftarrow \mathbf{w} + \alpha\Phi(\mathbf{x})$ and $\mathbf{w} \cdot \Phi(\mathbf{x})$ required.
  - Compute $\Phi(\mathbf{x})$ on-the-fly and parallelize!

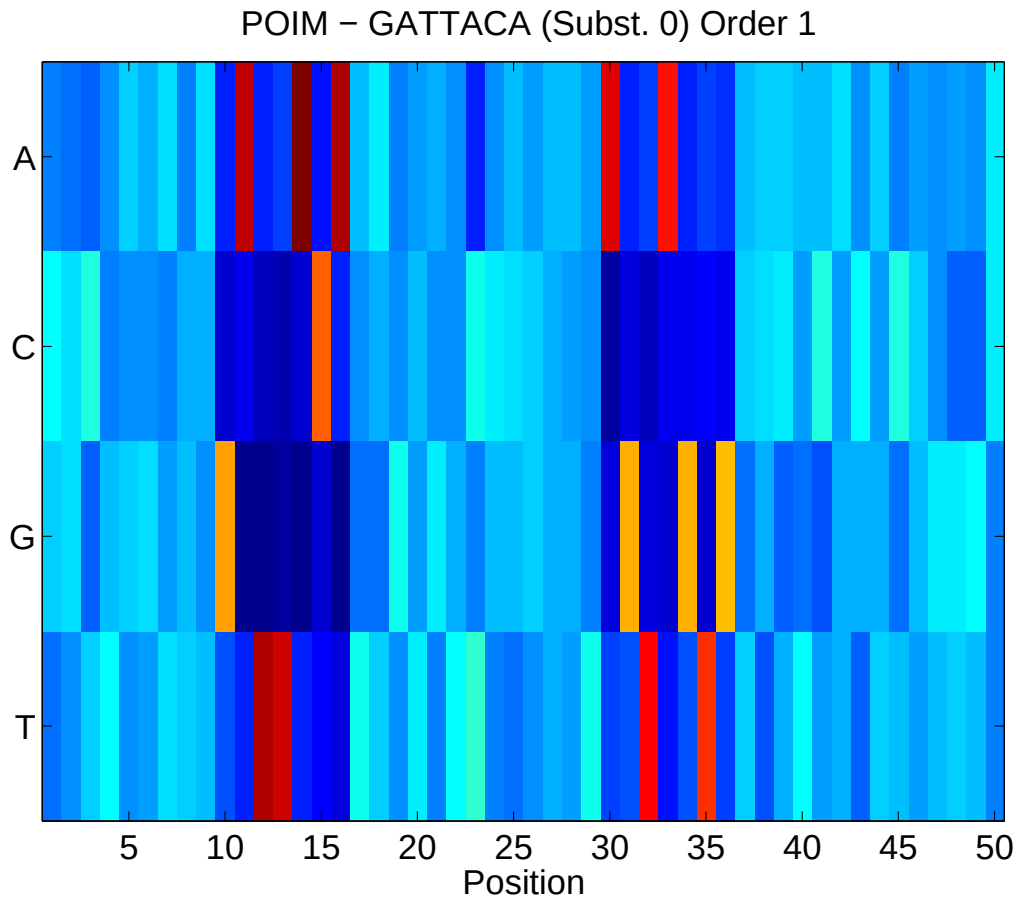
$\Rightarrow$ Ocas converges after 138 iterations in 41 hours.

Determining Gene Structures



Interpret Learned Results

Positional Oligomer Importance Matrices



$$Q(\mathbf{z}, j) \;:=\; \mathbb{E}\left[\, s(\mathbf{x}) \mid \mathbf{x}[j] = \mathbf{z} \,\right] - \mathbb{E}\left[\, s(\mathbf{x}) \,\right]$$

Project Subgoals