#### Templates :

**Section/Contribution Outline Template**

Section Title:

Probabilistic Inference and Generation of Immune Receptor Repertoires

Section Lead:

Yasamin Alemzadeh

Contributors:

**A. Purpose & Scope**

- What is the main question or focus of this section?

How can we accurately model the V(D)J recombination and somatic hypermutation (SHM) processes in adaptive immune receptors?

- Why is it important in the context of the review?

Understanding the probabilistic generation and diversification of BCRs and TCRs is crucial for interpreting high-throughput immune repertoire sequencing data. IGoR enables detailed modeling and statistical inference, which improves accuracy in identifying clonotypes and understanding immune dynamics, particularly in autoimmunity and disease states.

**B. Key Subsections & Points**

1. IGoR’s Core Probabilistic Framework:

* Bayesian network representation of recombination subprocesses
* Parameterization via conditional probabilities: gene usage, deletions, insertions
* Markov chain modeling for inserted nucleotides
* Model generality across TCRα, TCRβ, and BCR heavy chains

1. Scenario Enumeration and Likelihood Inference

* Smith-Waterman alignment to identify viable germline matches
* Scenario generation via decision tree traversal
* Computing scenario probability: 
* Calculating sequence generation probability  by marginalizing over scenarios

1. Learning and EM Algorithm

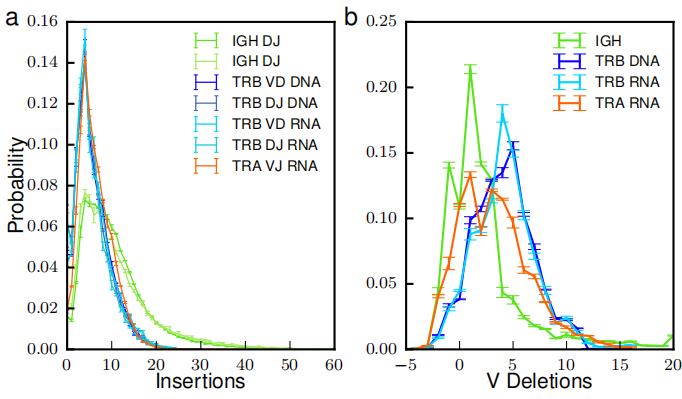
* Expectation-Maximization for model training
* Log-likelihood weighted by posterior probabilities
* Use of synthetic data for validation using Kullback-Leibler divergence
* Application on real repertoires (TCR and BCR)

1. Context-Aware Hypermutation Model

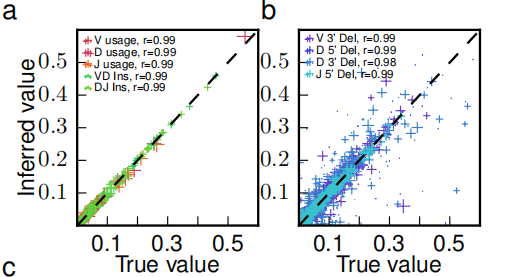
* PWM-based motif modeling for SHM hotspots
* Position-dependent and co-localized hypermutations
* Statistical validation using radial distribution function 

**C. Key References:**

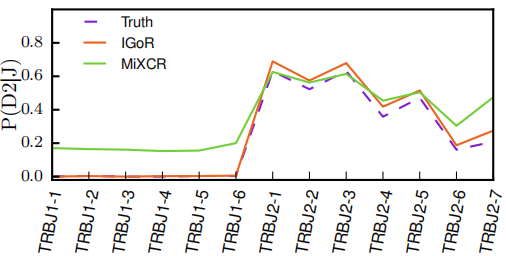
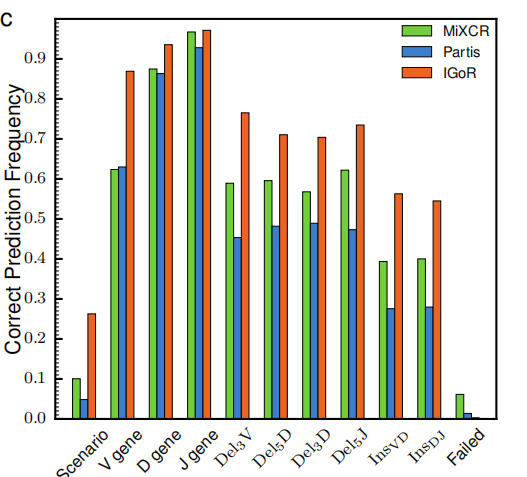
1. Marcou Q, Mora T, Walczak AM. High-throughput immune repertoire analysis with IGoR. Nat Commun. 2018.
2. Murugan A, Mora T, Walczak AM, Callan CG Jr. Statistical inference of T cell receptor generation probabilities. PNAS. 2012.
3. Ralph DK, Matsen FA. Consistency of VDJ Rearrangement and Hypermutation Models. PLoS Comput Biol. 2016.
4. Elhanati Y, Sethna Z, Marcou Q, Callan CG, Mora T, Walczak AM. Predicting the spectrum of TCR repertoire sharing. Phil Trans B. 2015.
5. **Planned Figures/Tables:**



Scenario multiplicity and probability distribution for reads



Repertoire-level distribution differences between IGoR and deterministic alignments



Accuracy comparison with MiXCR and Partis; capturing biological constraints (e.g., D-J pairs)

**E. Open Questions/Controversies**

* How well do PWM-based SHM models generalize to full-length reads and novel motifs?
* Can the computational cost of IGoR be reduced for large-scale clinical datasets?
* To what extent does IGoR’s probabilistic Pgen account for real antigen-driven selection vs convergent recombination?
* How transferable is the model across species (e.g., mouse to human)?

**Tool/Model Summary Template**

1. Tool/Model Name:

IGoR (Inference and Generation of Repertoires)

1. Reference(s):

Marcou Q, Mora T, Walczak AM. High-throughput immune repertoire analysis with IGoR. Nat Commun. 2018.

<https://www.nature.com/articles/s41467-017-01033-6>

3. Category:

Probabilistic generative model, Repertoire inference, SHM modeling

4. Main Purpose/Function:

To infer probabilistic recombination and hypermutation models from immune receptor data, enabling precise scenario enumeration, generative simulations, and likelihood estimation.

5. Key Features:

* Bayesian network modeling of V(D)J recombination
* Markov chain insertion model
* PWM-based SHM modeling
* ayesian network modeling of V(D)J recombination
* Markov chain insertion model
* PWM-based SHM modeling
* Support for BCR and TCR (α, β, heavy)

6. Input Requirements:

* Immune receptor sequences (nucleotide-level, e.g., FASTA)
* Germline segment databases (e.g., IMGT)
* User-specified model configuration (optional)

7. Output:

* Ranked recombination scenarios per sequence
* Generation probabilities (Pgen)
* Mutation profiles
* Synthetic repertoires

8. Validation/Evaluation:

* Benchmarking against synthetic datasets with known ground truth
* Quantitative comparison with MiXCR and Partis
* KL divergence to assess parameter recovery
* Real data application confirms learned dependency structures

9. Strengths:

* Probabilistic instead of heuristic assignment
* Accurately handles ambiguity in recombination scenarios
* Models SHM context explicitly
* Supports rare rearrangement patterns and reverse D segments

10. Limitations:

* High computational cost for large datasets
* Limited support for class switch recombination (CSR)
* No direct antigen-specificity modeling
* SHM model limited to short reads

11. Applications in Autoimmunity:

* Estimate Pgen to identify overrepresented clonotypes
* Disentangle SHM from convergent recombination
* Use in background generation for disease-association studies (e.g., SLE, MS)
* Detect abnormal rearrangements (e.g., tandem D segments)

1. Notable Citations/Use Cases:

* Marcou et al., 2018: SHM motif discovery in IGH repertoires
* Elhanati et al., 2015: Predicting TCR sharing across individuals
* Murugan et al., 2012: TCRβ dependency modeling (D-J)

13. Reviewer’s Comments:

IGoR stands out for its comprehensive probabilistic treatment of immune repertoire generation. Future improvements could involve GPU acceleration, long-read SHM modeling, and direct integration with selection modeling tools like SONIA.